Hierarchical models II

May 21, 2023

Overview

1. Normal model with exchangeable parameters

2. Example: parallel experiments in eight schools

Normal model with exchangeable parameters

- The observed data are normally distributed with a different mean for each 'group' or 'experiments', with known variance, and a normal population distribution for the group means.
- The hierarchical normal model sometimes could be termed the one-way normal random effects model with known data variance.
- Consider J independent experiments, with experiment j estimating the parameter θ_j from n_j independent normally distributed data points, y_{ij}, each with known error variance σ^2; that is,

$$y_{ij}|\theta_j \sim \mathrm{N}(\theta_j, \sigma^2), \text{ for } i = 1, \ldots, n_j; \;\; j = 1, \ldots, J.$$

Normal model data structure

- The mean of each group j is

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

- with sampling variance

$$\sigma_j^2 = \sigma^2 / n_j$$

- We can then write the likelihood for each θ_j using $\bar{y}_j$ instead of y_{ij}

$$\overline{y}_{.j} | \theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2),$$

The two likelihood functions about θ_j are equivalent. Please show the equivalence.

Constructing a prior distribution from pragmatic considerations

- Let us consider what sorts of posterior estimates might be reasonable for θ, given data (y_{ij})
- The pool estimate of overall mean is

$$\overline{y}_{..} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \overline{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}.$$

- The theoretical analysis of variance table is as follows, where τ^2 is the variance of $\theta_1, \dots, \theta_J$.

	df	SS	MS	$\mathrm{E}(\mathrm{MS}\|\sigma^2, \tau)$
Between groups	$J-1$	$\sum_i \sum_j (\overline{y}_{.j} - \overline{y}_{..})^2$	$\mathrm{SS}/(J-1)$	$n\tau^2 + \sigma^2$
Within groups	$J(n-1)$	$\sum_i \sum_j (y_{ij} - \overline{y}_{.j})^2$	$\mathrm{SS}/(J(n-1))$	σ^2
Total	$Jn-1$	$\sum_i \sum_j (y_{ij} - \overline{y}_{..})^2$	$\mathrm{SS}/(Jn-1)$	

The hierarchical model

- For the convenience of conjugacv, we assume that the parameter θ_j are drawn from a normal distribution with hyperparameters (μ, τ):

$$
\begin{aligned}
p(\theta_1, \ldots, \theta_J | \mu, \tau) &= \prod_{j=1}^{J} \mathrm{N}(\theta_j | \mu, \tau^2) \\
p(\theta_1, \ldots, \theta_J) &= \int \prod_{j=1}^{J} \left[\mathrm{N}(\theta_j | \mu, \tau^2)\right] p(\mu, \tau) d(\mu, \tau).
\end{aligned}
$$

- that is, the θ_js are conditionally independent given (μ, τ).
- We assign a noninformative uniform hyperprior distribution to μ, given τ:

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau).$$

The joint posterior distribution

- combining the sampling model for the observable y_{ij}s and the prior distribution yields the joint posterior distribution of all the parameters and hyperparameters, which we can express in terms of the sufficient statistics, $\bar{y}_j$:

$$
\begin{aligned}
p(\theta, \mu, \tau | y) &\propto p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta) \\
&\propto p(\mu, \tau) \prod_{j=1}^{J} \mathrm{N}(\theta_j | \mu, \tau^2) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j} | \theta_j, \sigma_j^2),
\end{aligned}
$$

- parameters σ_j are assumed known for this analysis.

The conditional posterior distribution

- The parameters θ_j are independent in the prior distribution given μ and τ and appear in different factors in the likelihood; thus, the conditional posterior distribution $p(\theta|\mu, \tau, y)$ factors into J components.
- Please deduct the posterior distribution. (Check the first order term and second order term of the joint posterior distribution with θ_j)

The conditional posterior distribution

- The conditional posterior distributions for the θ_js are independent, and

$$\theta_j|\mu, \tau, y \sim \mathrm{N}(\hat{\theta}_j, V_j),$$

where

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\overline{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}.$$

The marginal posterior distribution of the hyperparameters

- For the hierarchical model, we can simply consider the information supplied by the data about the hyperparameters directly

$$p(\mu, \tau | y) \propto p(\mu, \tau) p(y | \mu, \tau).$$

- Please deduct the likelihood $p(y|\mu, \tau)$. (Could be down through integration of $p(\bar{y}_j|\theta_j)p(\theta_j|\mu, \tau)$ with respect to θ_j)

The marginal posterior distribution of the hyperparameters

- The likelihood is normal

$$\overline{y}_{.j}|\mu, \tau \sim \mathrm{N}(\mu, \sigma_j^2 + \tau^2).$$

- The marginal posterior density is

$$p(\mu, \tau|y) \propto p(\mu, \tau) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j}|\mu, \sigma_j^2 + \tau^2).$$

The marginal posterior distribution of the hyperparameters

- With the joint posterior distribution of $p(\mu, \tau | y)$, we can find the corresponding density of $\mu | \tau, y$ combining the data with the uniform prior density $p(\mu | \tau)$ that

$$\mu | \tau, y \sim N(\hat{\mu}, V_\mu)$$

- Please deduct the posterior parameters (again, done by checking the first and second order terms of $p(\mu, \tau | y)$ with respect to μ)

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} \overline{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}} \quad \text{and} \quad V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}.$$

The marginal posterior distribution of the hyperparameters

- The posterior distribution of τ could be obtained analytically

$$\begin{aligned} p(\tau|y) &= \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \\ &\propto \frac{p(\tau)\prod_{j=1}^{J}\mathrm{N}(\overline{y}_{.j}|\mu,\sigma_j^2+\tau^2)}{\mathrm{N}(\mu|\hat{\mu},V_\mu)}. \end{aligned}$$

- If we set μ to $\hat{\mu}$, the expression would be

$$\begin{aligned} p(\tau|y) &\propto \frac{p(\tau)\prod_{j=1}^{J}\mathrm{N}(\overline{y}_{.j}|\hat{\mu},\sigma_j^2+\tau^2)}{\mathrm{N}(\hat{\mu}|\hat{\mu},V_\mu)} \\ &\propto p(\tau)V_\mu^{1/2}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left(-\frac{(\overline{y}_{.j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right), \end{aligned}$$

Posterior predictive distributions

We consider two scenarios:

- future data $\tilde{y}$ from the current set of batches, with means $\theta = (\theta_1, \ldots, \theta_J)$ which could be obtained from $p(\theta, \mu, \tau | y)$
- future data $\tilde{y}$ from $\tilde{J}$ future batches with means $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_J)$. For this one, we need to draw (μ, τ) from the posterior distribution, then draw $\tilde{J}$ new parameters $\tilde{\theta}$ from $p(\tilde{\theta}_j | \mu, \tau)$, then draw $\tilde{y}$ given $\tilde{\theta}$

Both method could be done through obtaining $\tilde{y}$ from

$$y_{ij} | \theta_j \sim \mathrm{N}(\theta_j, \sigma^2), \text{ for } i = 1, \ldots, n_j; \;\; j = 1, \ldots, J.$$

Example: parallel experiments in eight schools

- A study performed for the Educational Testing Service to analyze the effects of special coaching programs on test scores.
- Seperate random experiments were performed to estimate the effects of coaching programs for the SAT-V in each of eight high schools.
- The SAT are designed to be resistant to short-term efforts directed specifically toward improving performance on the test
- each of the eight schools in this study considered its short-term coaching program to be successful at increasing SAT scores.

Example: parallel experiments in eight schools

- The result is shown here

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Table 5.2

Example: parallel experiments in eight schools

- The estimated coaching effects we label y_j with sampling variance σ_j^2 play the same role in our model as $\bar{y}_j$ and σ_j^2 in previous study.
- the sample sizes in all of the eight experiments were relatively large.

Example: parallel experiments in eight schools

- **Separate estimates**: with separate estimate, treating each experiment separately and applying the simple normal analysis in each yields 95% posterior intervals that all overlap substantially.
- **A pooled estimate**: with pooled estimate, the overall mean is estimated as 7.7 with variance $(\sum_{j=1}^{8} \frac{1}{\sigma_j^2})^{-1} = 16.6$, which leads to the 95% posterior interval $[-0.5, 15.9]$
- Think about the question: would it be possible to have one school's observed effect be 28 just by chance, if the coaching effects in all eight schools were really the same?

Difficulties with the separate and pooled estimates

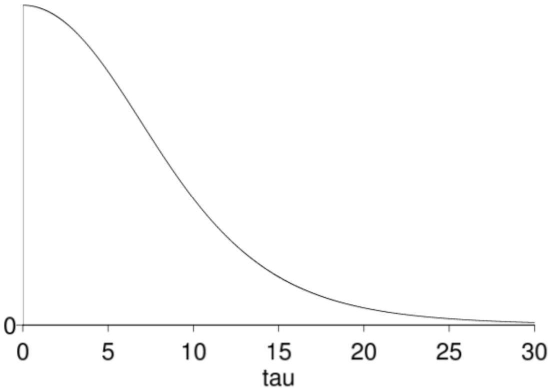
- The two extreme attitudes: the separate analyses consider each θ_j separately and the pooled estimate consider θ_1 with estimates 28 and standard error 15 comes from a normal distribution with mean 7.7 and standard deviation 4.1
- neither estimate is fully satisfactory, we would like a compromise that combines information from all eight experiments without assuming all the θ_js to be equal. The Bayesian analysis under the hierarchical model would be appropriate.

Posterior simulation under the hierarchical model

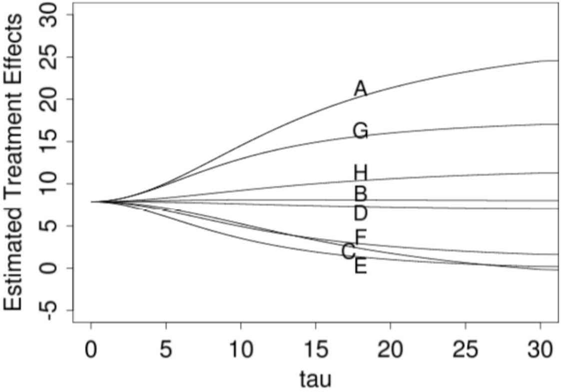
- Compute the posterior distribution of $\theta_1, \ldots, \theta_8$ based on the properties of normal model.
- Draw τ from posterior distribution $p(\tau|y)$ (MCMC), then draw μ from posterior distribution $p(\mu|\tau, y)$ (rnorm), finally draw θ_j from posterior distribution $p(\theta_j|\mu, \tau, y)$ (rnorm).
- The sampling standard deviations, σ_j, are assumed known and equal to the values in Table 5.2
- Assume the independent uniform prior densities on μ and τ

Example: Results

- Marginal posterior density function $p(\tau|y)$ is Figure 5.5, which shows $Pr(\tau > 25) \approx 0$

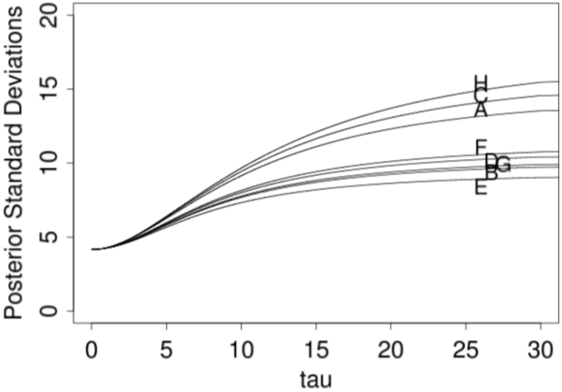


- Conditional posterior means $E(\theta_j|\tau, y)$ (averaging over μ) is Figure 5.6 as functions of τ, which shows as τ becomes larger, corresponding to more variability among schools, the estimates become more like the raw values in Table 5.2



Example: Results

- Conditional posterior standard deviation $sd(\theta_j|\tau, y)$ (standard deviation over μ) is Figure 5.7 as functions of τ, which shows as τ increase, the population distribution allows theieight effects to be more different from each other, when $\tau \to \infty$, the standard deviation approaching to the values in Table 5.2.



Example: Discussion

- Of substantial importance, we do not obtain an accurate summary of the data if we condition on the posterior mode of τ. The technique of conditioning on a modal value (for example, the maximum likelihood estimate) of a hyperparameter such as τ is often used in practice, but it ignores the uncertainty conveyed by the posterior distribution of the hyperparameter.