STA 360/602L: Module 5.3

Hierarchical normal models with constant variance: multiple groups

Dr. Olanrewaju Michael Akande

Comparing multiple groups

- Suppose we wish to investigate the mean (and distribution) of test scores for students at J different high schools.
- In each school j, where $j = 1, \ldots, J$, suppose we test a random sample of n_j students.
- Let y_{ij} be the test score for the ith student in school j, with $i = 1, \ldots, n_j$, with

$$y_{ij} | \theta_j, \sigma_j^2 \sim \mathcal{N}\left(\theta_j, \sigma_j^2\right)$$

  where for each school j, θ_j is the school-wide average test score, and σ_j^2 is the school-wide variance of individual test scores.
- This is what we did for the the Pygmalion study and job training data.

School testing example

- Option I: Classical inference for each school can be based on large sample 95% CI: $\bar{y}_j \pm 1.96\sqrt{s_j^2/n_j}$, where $\bar{y}_j$ is the sample average in school j, and s_j^2 is the sample variance in school j.
- Clearly, we can overfit the data within schools, for example, what if we only have 4 students from one of the schools? $\bar{y}_j$ can be a good estimate if n_j is large but it may be poor if n_j is small.
- Option II: alternatively, we might believe that $\theta_j = \mu$ for all j; that is, all schools have the same mean. This is the assumption (null hypothesis) in ANOVA models for example. We can also set $\sigma_j^2 = \sigma^2$ for all J.
- Option I ignores that the θ_j's should be reasonably similar, whereas option II ignores any differences between them.
- It would be nice to find a compromise! Borrowing information across, and shrinking our estimate towards a **grand mean** could be very useful here.

School testing example

- For the Pygmalion study and job training data, we focused on using priors that are independent between the groups.
- For example, in the conjugate case, we would have

$$\pi(\theta_j|\sigma_j^2) = \mathcal{N}\left(\mu_0, \frac{\sigma_j^2}{\kappa_0}\right)$$
$$\pi(\sigma_j^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

for some hyperparameters (constants), μ_0, κ_0, ν_0, and σ_0^2.

- In the semi-conjugate case,

$$\pi(\theta_j) = \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$
$$\pi(\sigma_j^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\gamma_0^2}{2}\right)$$

for some hyperparameters (constants), μ_0, σ_0^2, ν_0, and γ_0^2.

Hierarchical normal model

- Instead, we can assume that the θ_j's are drawn from a distribution based on the following: conceive of the schools themselves as being a random sample from all possible schools.

- For now, assume the variance is constant across schools. The hierarchical normal model assumes normal sampling models both within and between groups:

$$\begin{aligned} y_{ij}|\theta_j, \sigma^2 &\sim \mathcal{N}\left(\theta_j, \sigma^2\right); \quad i = 1, \ldots, n_j \\ \theta_j|\mu, \tau^2 &\sim \mathcal{N}\left(\mu, \tau^2\right); \quad j = 1, \ldots, J, \end{aligned}$$

  which gives us an extra level in the prior on the means, and leads to sharing of information across the groups in estimating the group-specific means.

- We have an extra variance parameter τ^2. Comparing τ^2 to σ^2 tells us how much of the variation in Y is due to within-group versus between-group variation.

Hierarchical normal model

- Standard semi-conjugate priors are given by

$$\pi(\mu) = \mathcal{N}\left(\mu_0, \gamma_0^2\right)$$
$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$
$$\pi(\tau^2) = \mathcal{IG}\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right).$$

with

  - μ_0: best guess of average of school averages
  - γ_0^2: set based on plausible ranges of values of μ
  - τ_0^2: best guess of variance of school averages
  - η_0: set based on how tight prior for τ^2 is around τ_0^2
  - σ_0^2: best guess of variance of individual test scores around respective school means
  - ν_0: set based on how tight prior for σ^2 is around σ_0^2.

Exchangeability

- This model relies heavily on exchangeability across units at each level.
- For example, we assume the schools are a random sample from the population of all schools, and the students within schools are a random sample of all the students in each school.
- This is not always completely true.
- Note: we can allow the variance to vary across schools if desired (and we will soon in fact).

Exchangeability

- Turns out that **conditional exchangeability** would be enough if we control for relevant variables in our modeling.
- For example, the schools in Chapel Hill/Carrboro are not entirely exchangeable.
- For example, Phoenix Academy is for students on long-term out-of-school suspension or who need to make up work due to extended absences (e.g., pregnancy), and Memorial Hospital School is for children battling serious illnesses.
- However, if we condition on school type (public, charter, private, special services, home), the schools may then be exchangeable.

Posterior inference

- Recall the model is

$$y_{ij}|\theta_j, \sigma^2 \sim \mathcal{N}\left(\theta_j, \sigma^2\right); \quad i = 1, \ldots, n_j$$
$$\theta_j|\mu, \tau^2 \sim \mathcal{N}\left(\mu, \tau^2\right); \quad j = 1, \ldots, J,$$

- Under our prior specification, we can factor the posterior as follows:

$$\begin{aligned}
\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2|Y) \propto \; & p(y|\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2) \\
& \times p(\theta_1, \ldots, \theta_J|\mu, \sigma^2, \tau^2) \\
& \times \pi(\mu, \sigma^2, \tau^2) \\
\\
= \; & p(y|\theta_1, \ldots, \theta_J, \sigma^2) \\
& \times p(\theta_1, \ldots, \theta_J|\mu, \tau^2) \\
& \times \pi(\mu) \cdot \pi(\sigma^2) \cdot \pi(\tau^2) \\
\\
= \; & \left\{ \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2) \right\} \\
& \times \left\{ \prod_{j=1}^{J} p(\theta_j|\mu, \tau^2) \right\} \\
& \times \pi(\mu) \cdot \pi(\sigma^2) \cdot \pi(\tau^2)
\end{aligned}$$

Full conditional for grand mean

- The full conditional distribution of μ is proportional to the part of the joint posterior $\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2 | Y)$ that involves μ.

- That is,

$$\pi(\mu | \theta_1, \ldots, \theta_J, \sigma^2, \tau^2, Y) \propto \left\{ \prod_{j=1}^{J} p(\theta_j | \mu, \tau^2) \right\} \cdot \pi(\mu).$$

- This looks like the full conditional distribution from the one-sample normal case, so you can show that

$$\pi(\mu | \theta_1, \ldots, \theta_J, \sigma^2, \tau^2, Y) = \mathcal{N}\left(\mu_n, \gamma_n^2\right) \quad \text{where}$$

$$\gamma_n^2 = \frac{1}{\dfrac{J}{\tau^2} + \dfrac{1}{\gamma_0^2}}; \qquad \mu_n = \gamma_n^2 \left[\frac{J}{\tau^2} \bar{\theta} + \frac{1}{\gamma_0^2} \mu_0 \right]$$

and $\bar{\theta} = \frac{1}{J} \sum\limits_{j=1}^{J} \theta_j$.

Full conditionals for group means

- Similarly, the full conditional distribution of each θ_j is proportional to the part of the joint posterior $\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2 | Y)$ that involves θ_j.

- That is,

$$\pi(\theta_j|\mu, \sigma^2, \tau^2, Y) \propto \left\{ \prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2) \right\} \cdot p(\theta_j|\mu, \tau^2)$$

- Those terms include a normal for θ_j multiplied by a product of normals in which θ_j is the mean, again mirroring the one-sample case, so you can show that

$$\pi(\theta_j|\mu, \sigma^2, \tau^2, Y) = \mathcal{N}\left(\theta_j^\star, \nu_j^\star\right) \quad \text{where}$$

$$\nu_j^\star = \frac{1}{\dfrac{n_j}{\sigma^2} + \dfrac{1}{\tau^2}}; \qquad \theta_j^\star = \nu_j^\star \left[\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\mu \right]$$

Full conditionals for group means

- Our estimate for each θ_j is a weighted average of $\bar{y}_j$ and μ, ensuring that we are borrowing information across all levels through μ and τ^2.

- The weights for the weighted average is determined by relative precisions from the data and from the second level model.

- The groups with smaller n_j have estimated $\theta_j^\star$ closer to μ than schools with larger n_j.

- Thus, degree of shrinkage of θ_j depends on ratio of within-group to between-group variances.

Full conditionals for across-group variance

- The full conditional distribution of τ^2 is proportional to the part of the joint posterior $\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2 | Y)$ that involves τ^2.
- That is,

$$\pi(\tau^2 | \theta_1, \ldots, \theta_J, \mu, \sigma^2, Y) \propto \left\{ \prod_{j=1}^{J} p(\theta_j | \mu, \tau^2) \right\} \cdot \pi(\tau^2)$$

- As in the case for μ, this looks like the one-sample normal problem, and our full conditional posterior is

$$\pi(\tau^2 | \theta_1, \ldots, \theta_J, \mu, \sigma^2, Y) = \mathcal{IG}\left(\frac{\eta_n}{2}, \frac{\eta_n \tau_n^2}{2} \right) \quad \text{where}$$

$$\eta_n = \eta_0 + J; \qquad \tau_n^2 = \frac{1}{\eta_n} \left[\eta_0 \tau_0^2 + \sum_{j=1}^{J} (\theta_j - \mu)^2 \right].$$

Full conditionals for within-group variance

- Finally, the full conditional distribution of σ^2 is proportional to the part of the joint posterior $\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2 | Y)$ that involves σ^2.
- That is,

$$\pi(\sigma^2|\theta_1, \ldots, \theta_J, \mu, \tau^2, Y) \propto \left\{ \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2) \right\} \cdot \pi(\sigma^2)$$

- We can again take advantage of the one-sample normal problem, so that our full conditional posterior is

$$\pi(\sigma^2|\theta_1, \ldots, \theta_J, \mu, \tau^2, Y) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right) \quad \text{where}$$

$$\nu_n = \nu_0 + \sum_{j=1}^{J} n_j; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0 \sigma_0^2 + \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2\right].$$

What's next?

Move on to the readings for the next module!