SDS 383D Exercises 4: Intro to Hierarchical Models

Tianqi Chen (tc34927)

1 Math tests

The data set in "mathtest.csv" shows the scores on a standardized math test from a sample of 10th-grade students at $P = 100$ different U.S. urban high schools, all having enrollment of at least 400 10th-grade students. (A lot of educational research involves "survey tests" of this sort, with tests administered to all students being the rare exception.)

Let θ_i be the underlying mean test score for school i, and let y_{ij} be the score for the jth student in school i. Our main parameter of interest is θ, the vector of school-level means. We'll assume that each student's test score is normally distributed around the corresponding school-level mean: $y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$ for $i = 1, \ldots, P$ and $j = 1, \ldots, N_i$.

(A) Show (somewhat trivially) that the maximum likelihood estimate for θ is just the vector of sample means: $\hat{\theta}_{\mathrm{MLE}} = (\bar{y}_1, \ldots, \bar{y}_P)$.

(B) Make a plot that illustrates the following fact: extreme school-level averages $\bar{y}_i$ (both high and low) tend to be at schools where fewer students were sampled. Explain briefly why this would be.

(C) Fit the following two-level hierarchical model to these data via Gibbs sampling:

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

As a starting point, use a flat prior on μ, Jeffreys' prior on σ^2, and an inverse-gamma(1/2, 1/2) prior on τ^2. Your Gibbs sampler should cycle between the complete conditional posterior distributions for each of these parameters in turn, as well as θ (the vector of means). While you could update each θ_i individually, I encourage you to think about it as a vector whose conditional distribution is multivariate normal, and whose covariance matrix happens to be diagonal. This view will generalize more readily to future problems.

(D) Express the conditional posterior mean for θ_i in the following form:

$$E(\theta_i \mid y, \tau^2, \sigma^2, \mu) = \kappa_i \mu + (1 - \kappa_i)\bar{y}_i \,,$$

i.e. a convex combination of prior mean and data mean. Here κ_i is a *shrinkage coefficient* whose form you should express in terms of the model hyperparameters. In the extreme case where κ_i is 1, then the data ($\bar{y}_i$) are essentially ignored, and the posterior mean is "shrunk" all the way back to the prior mean. In the other extreme where κ_i is 0, the prior mean is ignored, and the posterior mean is entirely "un-shrunk" compared to the MLE for θ_i.

For each draw of your MCMC, calculate κ_i for each school, and save the posterior draws. Average these MCMC samples to calculate $\bar{\kappa}_i$, the posterior mean of this shrinkage coefficient. Plot $\bar{\kappa}_i$ for each school as a function of that school's sample size, and comment on what you see.

(E) Observe that an equivalent way to write your model involves the following decomposition:

$$y_{ij} = \mu + \delta_i + e_{ij}$$

where $\delta_i \sim N(0, \tau^2\sigma^2)$ and $e_{ij} \sim N(0, \sigma^2)$. (In the paper by Gelman that I've asked you to read, he writes it this way, where the school-level "offsets" are centered at zero, although he doesn't scale these offsets by σ the way I prefer to do.) To translate between the two parameterizations, just observe that in the previous version, $\theta_i = \mu + \delta_i$.

Conditional on the "grand mean" μ, but *marginally* over both δ_i and e_{ij}, compute the following two covariances:

- $\mathrm{cov}(y_{i,j}, y_{i,k})$, $j \neq k$
- $\mathrm{cov}(y_{i,j}, y_{i',k})$, $i \neq i'$ and $j \neq k$

Does this make sense to you? Why or why not?

(F) Does the assumption that σ^2 is common to all schools look justified in light of the data?

Answer

Write down the joint p.d.f.

$$p\left(\mathbf{y}, \Theta\right) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{P}\sum_{j=1}^{N_i}(y_{ij}-\theta_i)^2\right)\exp\left(-\frac{1}{2(\tau\sigma)^2}\sum_{i=1}^{P}(\theta_i-\mu)^2\right)$$

where $\mathbf{y} = \{y_{ij} : 1 \leq i \leq P,\ 1 \leq j \leq N_i\}$, $\Theta = (\theta_1, \ldots, \theta_P)$.

(A) The log-likelihood up to addition of some constant is

$$\begin{aligned} l(\Theta) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{P}\sum_{j=1}^{N_i}(y_{ij}-\theta_i)^2 \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{P} N_i(\theta_i - \bar{y}_i)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{P} N_i S_{y_i} \end{aligned}$$

where $\bar{y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i$ and $S_{y_i} = \sum_{i=1}^{N_i}(y_{ij} - \bar{y}_i)^2/N_i$.

Therefore, to maximize the likelihood, it is equivalent to maximize the log-likelihood, which can be expressed as a sum of quadratic functions, each of which

the estimator $\hat{\Theta}_{\mathrm{MLE}}$ should be $(\bar{y}_1, \ldots, \bar{y}_P)$

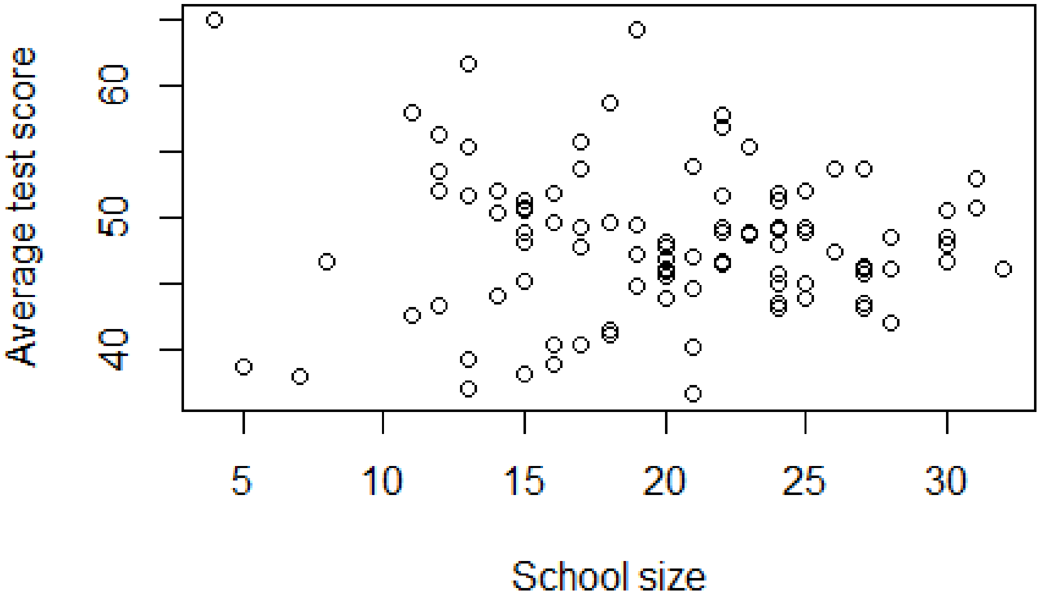

Figure 1: Average test score vs School sample size

(B) A reason why extreme average scores tend to associate with schools of few samples could be that averages by fewer samples are more likely to have larger variance and thus extreme values. Another reason might be that some of the schools with fewer sample size also provide above/below average level of math education.

(C) Let $\kappa = 1/\tau^2$ and $s = \sigma^2$. According to the model specification

$$\begin{aligned}
p(s) &\propto s^{-1} \\
p(\kappa) &\propto \kappa^{\frac{1}{2}-1} \exp\left(-\frac{\kappa}{2}\right) \\
p(\mu) &\propto 1 \\
p(\Theta|\mu, s) &\propto \kappa^{\frac{P}{2}} s^{-\frac{P}{2}} \exp\left(-\frac{\kappa}{2s}\sum_{i=1}^{P}(\theta_i - \mu)^2\right) \\
p(\mathbf{y}|\Theta, s) &\propto s^{-\frac{N}{2}} \exp\left(-\frac{1}{2s}\sum_{i=1}^{P}\sum_{j=1}^{N_i}(y_{ij} - \theta_i)^2\right)
\end{aligned}$$

The posterior distribution can be written as

$$p(\Theta, s, \kappa, \mu|\mathbf{y}) \propto s^{-\frac{N+P}{2}-1}\kappa^{\frac{P+1}{2}-1} \exp\left(-\frac{\kappa}{2}\right) \exp\left(-\frac{1}{2s}\left(\sum_{i=1}^{P}\sum_{j=1}^{N_i}(y_{ij} - \theta_i)^2 + \kappa\sum_{i=1}^{P}(\theta_i - \mu)^2\right)\right)$$

The complete conditionals of posterior are

$$\begin{aligned}
s|\Theta, \kappa, \mu &\propto \text{Inv-Gamma}\left(\frac{N+P}{2}, \frac{(N+P)V_s}{2}\right) \\
\kappa|\Theta, s, \mu &\propto \Gamma\left(\frac{P+1}{2}, \frac{(P+1)V_\kappa}{2}\right) \\
\mu|\Theta, s, \kappa &\propto \mathrm{N}\left(\bar{\theta}, s(\kappa P)^{-1}\right) \\
\theta_i|s, \kappa, \mu &\propto \mathrm{N}\left(V_{\theta_i}^{-1} m_i, sV_{\theta_i}^{-1}\right)
\end{aligned}$$

where

$$\begin{aligned}
N &= \sum_{i=1}^{P} N_i \\
V_s &= \frac{\sum_{i=1}^{P}\sum_{j=1}^{N_i}(y_{ij} - \theta_i)^2 + P\kappa S_\theta}{N+P} \\
S_\theta &= \frac{\sum_{i=1}^{P}(\theta_i - \mu)^2}{P} \\
V_\kappa &= \frac{PS_\theta/s + 1}{P+1} \\
\bar{\theta} &= \frac{\sum_{i=1}^{P}\theta_i}{P} \\
m_i &= \kappa\mu + N_i\bar{y}_i \\
V_{\theta_i} &= \kappa + N_i
\end{aligned}$$

(D) In part (C), we derived mean of $\theta_i|\mathbf{y},\kappa,s,\mu$, which is $\frac{\kappa\mu+N_i\bar{y}_i}{\kappa+N_i} = \frac{\frac{1}{\tau^2}\mu+N_i\bar{y}_i}{\frac{1}{\tau^2}+N_i}$. Therefore, $k_i = \frac{1}{1+\tau^2 N_i}$.

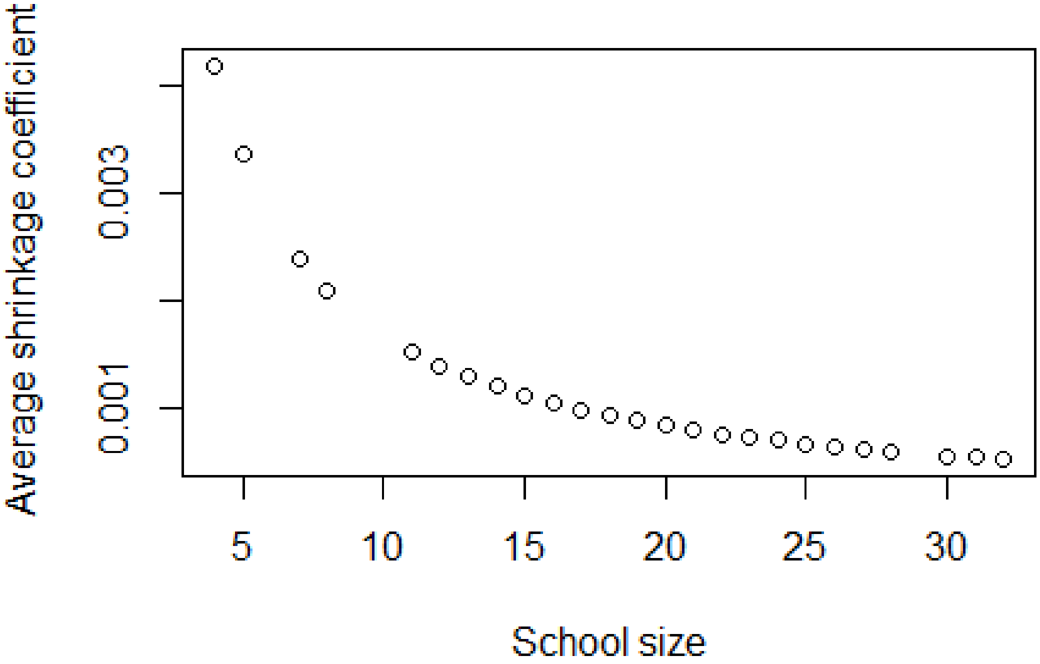

Figure 2: Average shrinkage coefficient vs School sample size

(E) The expectation of y_{ij} over δ_i and e_{ij} is

$$E_{\delta_i,e_{ij}}[y_{ij}] = \mu$$

By applying the law of total covariance, we have

$$\begin{aligned}\mathrm{cov}(y_{ij},y_{ik}) &= E_{\delta_i}\mathrm{cov}(y_{ij},y_{ik}|\delta_i) + \mathrm{cov}(E[y_{ij}|\delta_i],E[y_{ik}|\delta_i]) \\ &= E_{\delta_i}\mathrm{cov}(e_{ij},e_{ik}) + cov(\delta_i,\delta_i) \\ &= 0 + \tau^2\sigma^2 = \tau^2\sigma^2\end{aligned}$$

$$\begin{aligned}\text{cov}(y_{ij}, y_{i'k}) &= E_{\delta_i,\delta_{i'}}\text{cov}(y_{ij}, y_{i'k}|\delta_i, \delta_{i'}) + \text{cov}(E[y_{ij}|\delta_i], E[y_{i'k}|\delta_{i'}]) \\ &= E_{\delta_i}\text{cov}(e_{ij}, e_{i'k}) + cov(\delta_i, \delta_{i'}) \\ &= 0\end{aligned}$$

Yes, it makes sense.

(F) No.

2 Blood pressure

The data set in “bloodpressure.csv” contains data on repeated blood pressure measurements on 20 different subjects, ten of whom received a control medication (treatment=1) and ten of whom received an experimental medication (treatment = 2). Patients were randomized to receive the two treatments. The columns are the data of measurement, a numerical ID for the subject, the blood pressure measurement (systolic), and which treatment the patient received.

(A) Is the experimental medication effective at reducing blood pressure? Do the naive thing and perform a t-test for a difference of means, pooling all the data from treatment 1 into group 1, and all the data from treatment 2 into group 2. What does this t-test say about the difference between these two group means, and the standard error for the difference? Why is the t-test (badly) wrong?

(B) Now do something better, but still less than ideal. Calculate $\bar{y}_i$, the mean blood pressure measurement for each patient. Now treat each person-level mean as if it were just a single data point, and conduct a different t-test for mean blood pressure between treatment 1 and treatment 2. (If you're doing this correctly, you should have only ten ”observations” in each group, where each observation is actually a person-level mean.) What does this t-test say about the difference between these two group means, and the standard error for the difference? Why is the standard error so much bigger, and why is this appropriate? Even so, why is this approach (subtly) wrong?

(C) Now fit a two-level hierarchical model to this data, of the following form:

$$\begin{aligned}(y_{ij} \mid \theta_i, \sigma^2) &\sim \text{N}(\theta_i, \sigma^2) \\ (\theta_i \mid \tau^2, \sigma^2) &\sim \text{N}(\mu + \beta x_i, \tau^2\sigma^2)\end{aligned}$$

where y_{ij} is blood pressure measurement j on person i, and x_i is a dummy (0/1) variable indicating whether a patient received treatment 2, the experimental medication. Apply what you learned on the previous problem about sampling, hyperparameters, etc, but account for the extra wrinkle here, i.e. the presence of the βx_i term that shifts the mean between the treatment and control groups.

Write our your model's complete conditional distributions, and fit it. Make a histogram of the posterior distribution for β, which represents the treatment effect here. In particular, what are the posterior mean and standard deviation of β? How do these compare to the estimates and standard errors from the approaches in (A) and (B)?

(D) Your two-level model assumes that, conditional on θ_i, the y_{ij} are independent. Written concisely: $(y_{ij} \perp y_{ik} \mid \theta_i)$ for $j \neq k$.

There are many ways this assumption could break down. So check! Does this assumption look (approximately) sensible in light of the data? Provide evidence one way or another.

Answer

(A) The result by carrying out a Welch two-sample t-test seems to provide evidence that the experimental medication is effective. However, the naive t-test here is problematic because each subject is associated with multiple records.

(B)

(C) Let $\kappa = 1/\tau^2$, $s = \sigma^2$. Similar to part C in Section 1, the posterior distribution can be written as

$$p(\Theta, s, \kappa, \mu, \beta|\mathbf{y}) \propto s^{-\frac{N+P}{2}-1} \kappa^{\frac{P+1}{2}-1} \exp\left(-\frac{\kappa}{2}\right) \exp\left(-\frac{1}{2s}\left(\sum_{i=1}^{P}\sum_{j=1}^{N_i}(y_{ij}-\theta_i)^2 + \kappa\sum_{i=1}^{P}(\theta_i - \mu - \beta x_i)^2\right)\right)$$

The complete conditionals of posterior are

$$\begin{aligned}
s|\Theta, \kappa, \mu, \beta &\propto \text{Inv-Gamma}\left(\frac{N+P}{2}, \frac{(N+P)V_s}{2}\right) \\
\kappa|\Theta, s, \mu, \beta &\propto \Gamma\left(\frac{P+1}{2}, \frac{(P+1)V_\kappa}{2}\right) \\
\mu|\Theta, s, \kappa, \beta &\propto \mathrm{N}\left(\bar{\theta} - \beta\bar{x}, s(\kappa P)^{-1}\right) \\
\beta|\Theta, s, \kappa, \mu &\propto \mathrm{N}\left(x^T(\theta - \mu)/\|x\|^2, s\left(\kappa\sum_{i=1}^{P}\|x\|^2\right)^{-1}\right) \\
\theta_i|s, \kappa, \mu, \beta &\propto \mathrm{N}\left(V_{\theta_i}^{-1}m_i, sV_{\theta_i}^{-1}\right)
\end{aligned}$$

where

$$
\begin{aligned}
N &= \sum_{i=1}^{P} N_i \\
V_s &= \frac{\sum_{i=1}^{P}\sum_{j=1}^{N_i}(y_{ij} - \theta_i - \beta x_i)^2 + P\kappa S_\theta}{N + P} \\
S_\theta &= \frac{\sum_{i=1}^{P}(\theta_i - \mu - \beta x_i)^2}{P} \\
V_\kappa &= \frac{PS_\theta/s + 1}{P + 1} \\
\bar{\theta} &= \frac{\sum_{i=1}^{P}\theta_i}{P} \\
m_i &= \kappa(\mu + \beta x_i) + N_i\bar{y}_i \\
V_{\theta_i} &= \kappa + N_i
\end{aligned}
$$

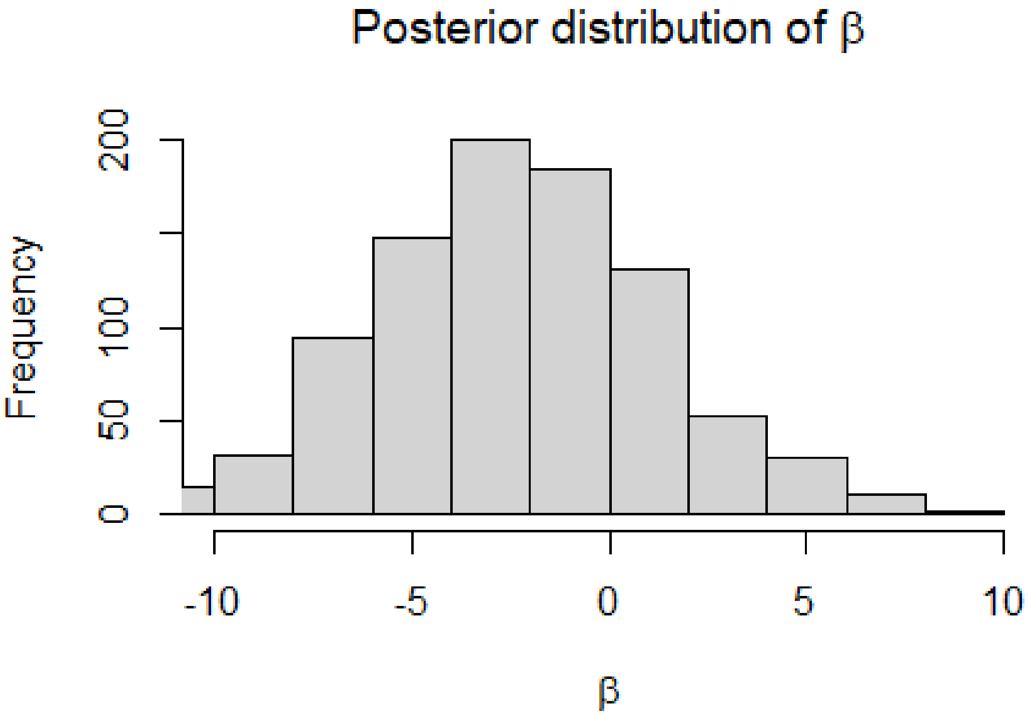

Figure 3: Posterior β distribution by Gibbs sampler

(D)

(E)