Ch 4 Hierarchical models: math test data analysis

Friday, March 29, 2019 1:29 AM

Consider the dataset in "mathtest.csv", which shows the scores on a standardized math test from a sample of 10th grade students at 100 US urban high schools. Let θ_i be the underlying mean test score for school i, and let y_{ij} be the score for the jth student at school i. Notice the extreme school-level averages $\bar{y}_i$ (both high and low) tend to be at schools where fewer students were sampled.

1.) Explain briefly why this would be.

The sample variance is $\frac{\sigma^2}{n}$ where σ^2 is the true population variance. Therefore, it is easy to see that when n is small, the variance is larger. The variance as a function of sample size is a funnel plot.

The mean value theorem states that the sample mean would converge to the population mean for large n. That is, as the sample size grows, the probability that the sample mean is very different from the true population mean goes to 0. Now, we do not know what the true value of θ_i are for the schools that had the fewest samples but it is not unreasonable to assume that the variance of θ_i is not too large. Therefore, supposing that the θ_i are not too far from their overall mean μ, the $\bar{y}_i$ that come from schools with large sample sizes would also be relatively close in value to their θ_i and μ by the mvt. However, when the sample size is small, the mvt no longer holds and $\bar{y}_i$ can deviate greatly from θ_i, and by extension, $\bar{y}_{j \neq i}$.

2.) Fit this normal hierarchical model to the math test data via Gibbs sampling:

$$y_{ij} \mid \theta_i \sim N(\theta_i, \sigma^2)$$

$$\theta_i \sim N(\mu, \tau^2 \sigma^2)$$

$$E(y) = E(E(y \mid \theta_i)) = \mu$$

Decide upon sensible priors for the unknown model parameters (μ, σ^2, τ^2).

In order to implement gibbs sampling, we need to derive the full conditionals $p(\theta_i \mid \theta_{-i}, \text{data})$. First I will state the pieces of the joint posterior. To make the derivations easier, I will swith to working with precisions. Let $\lambda = \frac{1}{\sigma^2}$, $\nu = \frac{1}{\tau^2}$

$$p(y_{ij} \mid \theta_i, \lambda) \sim N(\theta_i, \tfrac{1}{\lambda})$$

$$p(\theta_i \mid \mu, \nu, \lambda) \sim N(\mu, \tfrac{1}{\lambda \nu})$$

I will assign conjugate priors to μ, λ, and ν. Let

$$p(\mu \mid 0, \gamma) \sim N(0, \tfrac{1}{\gamma})$$

$$p(\lambda \mid a, b) \sim Ga(a, b)$$

$$p(\nu \mid c, d) \sim Ga(c, d)$$

We can now construct the full posterior

$$p(\underset{\sim}{\theta}, \lambda, \mu, \nu \mid \underset{\sim}{y}_{ij}) \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} [p(y_{ij} \mid \theta_i, \lambda)] \prod_{i=1}^{n} [p(\theta_i \mid \mu, \nu, \lambda)] p(\mu) p(\nu) p(\lambda)$$

$$p(\underset{\sim}{\theta},\lambda,\mu,\nu \mid \underset{\sim}{y_{ij}}) \propto \prod_{i=1}^{n}\prod_{j=1}^{m_i}\left[p(y_{ij}\mid\theta_i,\lambda)\right]\prod_{i=1}^{n}\left[p(\theta_i\mid\mu,\nu,\lambda)\right]p(\mu)\,p(\nu)\,p(\lambda)$$

$$\propto \prod_{i=1}^{n}\prod_{j=1}^{m_i}\lambda^{1/2}\exp\left(-\frac{\lambda}{2}(y_{ij}-\theta_i)^2\right)\prod_{i=1}^{n}\left[(\nu\lambda)^{1/2}\exp\left(-\frac{\nu\lambda}{2}(\theta_i-\mu)^2\right)\right]\exp\left(-\frac{\gamma}{2}(\mu-\mu_0)^2\right)\lambda^{a-1}\exp(-b\lambda)\;\nu^{c-1}\exp(-d\nu)$$

$$=\lambda^{\frac{\sum_{i=1}^{n}m_i}{2}}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\theta_i)^2\right)(\nu\lambda)^{\frac{n}{2}}\exp\left(-\frac{\nu\lambda}{2}\sum_{i=1}^{n}(\theta_i-\mu)^2\right)\exp\left(-\frac{\gamma}{2}(\mu-\mu_0)^2\right)\lambda^{(a-1)}\exp(-b\lambda)\,\nu^{(c-1)}\exp(-d\nu)$$

Then the full conditionals are of the form

$$p(\theta_i\mid \underset{\sim}{y_i},\lambda,\mu,\nu) \propto \exp\left(-\frac{\lambda}{2}\sum_{j=1}^{m}(y_{ij}-\theta_i)^2\right)\exp\left(-\frac{\nu\lambda}{2}(\theta_i-\mu)^2\right)$$

$$=\exp\left(-\frac{\lambda}{2}\sum_{j=1}^{m}y_{ij}^2-2\theta_i y_{ij}+\theta_i^2\right)\exp\left(-\frac{\nu\lambda}{2}(\theta_i^2-2\theta_i\mu+\mu^2)\right)$$

$$\propto \exp\left(-\frac{\lambda}{2}\left(m\theta_i^2-2\theta_i\left(\sum_{j=1}^{m}y_{ij}\right)\right)-\frac{\nu\lambda}{2}\left(\theta_i^2-2\theta_i\mu\right)\right)$$

$$=\exp\left(-\frac{1}{2}\left[(m\lambda+\nu\lambda)\theta_i^2-2\theta_i\left(\lambda\sum_{j=1}^{m}y_{ij}+\nu\lambda\mu\right)\right]\right)$$

$$\propto \exp\left(-\frac{m\lambda+\nu\lambda}{2}\left(\theta_i-\frac{\lambda\sum_{j=1}^{m}y_{ij}+\nu\lambda\mu}{m\lambda+\nu\lambda}\right)^2\right)$$

$$\sim N\left(\frac{\lambda\sum_{j=1}^{m}y_{ij}+\nu\lambda\mu}{m\lambda+\nu\lambda},\ \frac{1}{m\lambda+\nu\lambda}\right)$$

$$p(\lambda\mid \underset{\sim}{y_{ij}},\mu,\underset{\sim}{\theta_i},\nu)\propto \lambda^{\left(\frac{\sum_{i=1}^{n}m_i+n}{2}+a-1\right)}\exp\left(-\lambda\left(\frac{1}{2}\sum\sum(y_{ij}-\theta_i)^2+\frac{\nu}{2}\sum_{i=1}^{n}(\theta_i-\mu)^2+b\right)\right)$$

$$\sim Ga\left(\frac{\sum_{i=1}^{n}m_i+n}{2}+a,\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\theta_i)^2+\frac{\nu}{2}\sum_{i=1}^{n}(\theta_i-\mu)^2+b\right)$$

$$p(\nu\mid\theta_i,\mu,c,d)\propto \nu^{\left(\frac{n}{2}+c-1\right)}\exp\left(-\nu\left(\frac{\lambda}{2}\sum_{i=1}^{n}(\theta_i-\mu)^2+d\right)\right)$$

$$\sim Ga\left(\frac{n}{2}+c,\ \frac{\lambda}{2}\sum_{i=1}^{n}(\theta_i-\mu)^2+d\right)$$

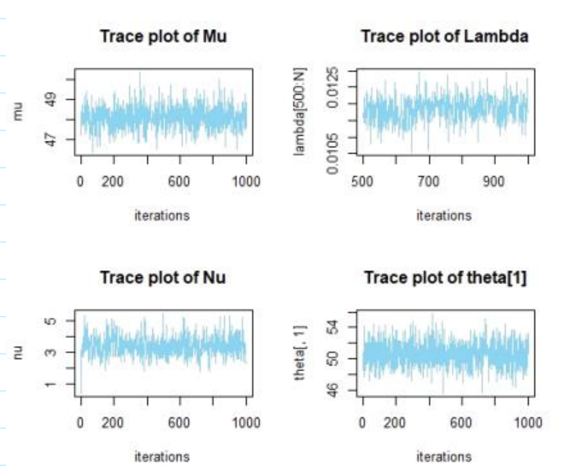
$$p(\mu\mid\theta_i,\gamma)\propto \exp\left(-\frac{\nu\lambda}{2}\sum_{i=1}^{n}(\theta_i-\mu)^2\right)\exp\left(-\frac{\gamma}{2}(\mu-\mu_0)^2\right)$$

$$\propto \exp\left(-\frac{\nu\lambda}{2}\left(n\mu^2-2\mu\sum_{i=1}^{n}\theta_i\right)-\frac{\gamma}{2}(\mu^2-2\mu\mu_0+\mu_0^2)\right)$$

$$\propto \exp\left(-\frac{\nu\lambda n+\gamma}{2}\left(\mu-\frac{\nu\lambda\sum_{i=1}^{n}\theta_i+\gamma\mu_0}{\nu\lambda n+\gamma}\right)^2\right)$$

$$\sim N\left(\frac{\nu\lambda\sum_{i=1}^{n}\theta_i+\gamma\mu_0}{\nu\lambda n+\gamma},\ \frac{1}{\nu\lambda n+\gamma}\right)$$

For the hyperparameters, I will let $\mu_0 = \text{average}(\bar{y}_i)$. I will set γ to something small so that the variance is large. This will make the prior of μ more uninformative. I will let $a=b=c=d=1$.



The posterior mean of $\mu = 48.119$

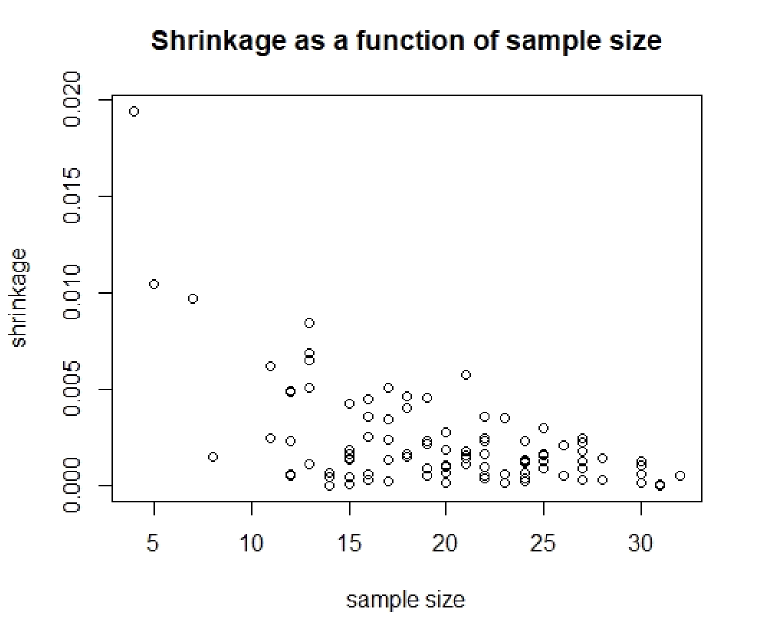
$\lambda = 0.012$

$\nu = 3.410$

3.) Define the shrinkage coefficient K_i as

$$K_i = \frac{\bar{y}_i - \hat{\theta}_i}{\bar{y}_i}$$

which tells you how much the posterior mean shrinks the observed sample mean. Plot this shrinkage coefficient (in absolute value) for each school as a function of that school's sample size, and comment.



The shrinkage values are higher for schools with the smaller samples. This is in line with our expectations. With fewer samples, the $\sum_{j=1}^{m} y_{ij}$ would be smaller and so the posterior mean of θ_i would be pulled towards the prior mean μ. This is the effect we want for this dataset. Since having fewer samples makes the sampling distribution more variable, we want to skrink these means towards the overall mean to remove some of the unwarraneed variance.