STA 630: Bayesian Inference - Chapter 6

Zeya Wang

University of Kentucky

Spring 2025

Lecture 11-12: Hierarchical Models

- Hierarchical normal models
- Math scores example
- Classical estimation
- Bayesian estimation
- Model Extensions - unequal variances
- Shrinkage effects
- Model Parameterizations

Introduction

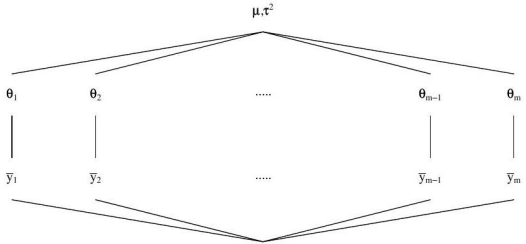
- The exercise of specifying a model over several levels is called **hierarchical modeling**, with each new distribution forming a new level in the hierarchy.
- In a hierarchical model, the observations are given distributions conditional on parameters, and the parameters in turn have distributions conditional on additional parameters called hyperparameters.
- Non-hierarchical models are usually inappropriate for hierarchical data
  - with few parameters, they generally cannot fit the data adequately
  - with many parameters, they tend to "overfit" the data.
- Hierarchical models allow information to be shared across groups of observations.

Hierarchical normal model

The hierarchical normal model is used to describe the heterogeneity of means across several populations.

$$\begin{aligned} Y_{ij} &\sim \mathcal{N}\left(\theta_j, \sigma^2\right) && \text{within-group model} \\ \theta_j &\sim \mathcal{N}\left(\mu, \tau^2\right) && \text{between-group model} \end{aligned}$$

The within-group sampling variability σ^2 is assumed to be constant across groups (we can eliminate this assumption).



Hierarchical normal model

The unknown quantities are:

- the group-specific means $\{\theta_1, \ldots, \theta_m\}$
- the within-group sampling variability σ^2
- the mean and variance of the population of group-specific means $\left(\mu, \tau^2\right)$

Posterior inference for these parameters can be made by constructing a Gibbs sampler which approximates $p\left(\theta_1, \ldots, \theta_m, \sigma^2, \mu, \tau^2 \mid y_1, \ldots, y_m\right)$ by iteratively sampling each parameter from its full conditional distribution.

Example: High school math achievement

- Representative sample of U.S. public schools (160 schools)
- Within each school, a random sample of students is selected (14 to 67) with a total of 7185 students.
- The primary outcome, Y_{ij}, is a standardized measure of math achievement for student i in school j.
- Additional covariates are collected on each student.

One-way ANOVA

In the classical one-way analysis of variance model:

$$Y_{ij} = \theta_j + \varepsilon_{ij} \quad \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

where σ^2 measures how much variation each individual student deviates from the school mean θ_j.

- The θ_j 's are referred to as fixed effects.
- The interest is in drawing inference for the individual means or the differences in means for the m groups.
- The hypothesis tests in ANOVA are:

  $H_0 : \theta_1 = \theta_2 = \ldots = \theta_{160}$ (same mean math scores across schools)

  H_1 : means are not all equal

One-way ANOVA

The classical approach uses a one-way analysis of variance (ANOVA) model.

	df	SS	MS	F
between groups	$m-1$	$\sum_{j=1}^{m} n_j \left(\bar{Y}_j - \bar{Y}\right)^2$	$SS/(m-1)$	$\frac{MSB}{MSW}$
within groups	$\sum_{j=1}^{m} n_j - m$	$\sum_{j=1}^{m}\sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_j\right)^2$	$SS/\left(\sum_j n_j - m\right)$	
total	$\sum_{j=1}^{m} n_j - 1$	$\sum_{j=1}^{m}\sum_{i} \left(Y_{ij} - \bar{Y}\right)^2$	$SS/\left(\sum_j n_j - 1\right)$	

where $n = \sum_{j=1}^{m} n_j, \quad \bar{Y}_j = \frac{1}{n_j}\sum_{j=1}^{m} Y_{ij}, \quad \bar{Y} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} Y_{ij}$

One-way ANOVA

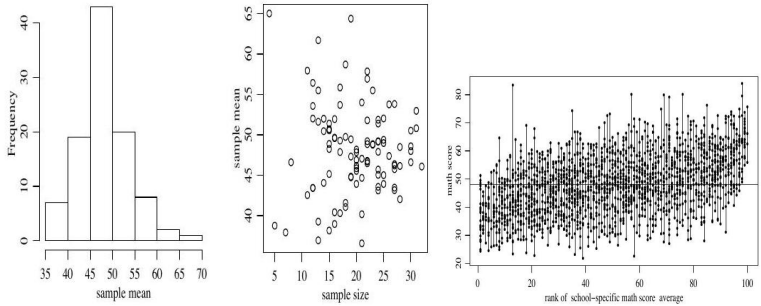
- If ratio of between to within mean squares is significantly < 1 (F-test p-value ≤ 0.05), reject H_0 and use separate estimates: $\hat{\theta}_j = \bar{Y}_j$ for each j.
- If the ratio of mean squares is not "statistically significant" (F-test p-value > 0.05), pooling the means is reasonable: $\hat{\theta}_1 = \ldots = \hat{\theta}_m = \bar{Y}$.

```
anova(aov(y ~ school))
Analysis of Variance Table
    Df Sum Sq Mean Sq F value Pr(>F)
school 159 64907 408 10.429 2.2e-16 ***
Residuals 7025 274970 39
```

So we either treat all the means as equal or all different. Two extremes.

One-way ANOVA

Let us look at the mean math scores and number of students across the 160 schools. Some schools perform better than others. There is also variability among students within each school



How much do US high schools vary in their mean mathematics achievement?

Random Effect Model

- The student level model is given by

$$Y_{ij} = \theta_j + \varepsilon_{ij} \quad \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

  where σ^2 measures how much variation each individual student deviates from the school mean θ_j.

- In addition, the high schools vary in their mean mathematics achievement

$$\theta_j \overset{iid}{\sim} \mathcal{N}\left(\mu, \tau^2\right)$$

In the second-level, the school-level means are viewed as random effects arising from a normal population, i.e., the school level means vary about an overall mean μ with variance τ^2.

Random Effect Model

- μ is the overall population mean, a fixed effect
- σ^2 is the within-group variance or variance component
- τ^2 is the between-group variance (it captures how much US high schools vary in their mean mathematics achievement)

There are 2 additional parameters versus the $m + 1$ in the fixed effects model.

Marginal model

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

$$Y_{ij} \sim \mathcal{N}\left(\mu, \tau^2 + \sigma^2\right)$$

where

$$\begin{aligned} \text{Cov}\left(Y_{ij}, Y_{i'j}\right) &= \tau^2 \\ \text{Cov}\left(Y_{ij}, Y_{i'j'}\right) &= 0 \quad \text{for } j \neq j' \end{aligned}$$

This model implies students within schools are exchangeable and student achievements across different schools are independent given the school effect.

Intraclass Correlation

The intraclass correlation

$$\text{Corr}\left(Y_{ij}, Y_{i'j}\right) = \frac{\tau^2}{\tau^2 + \sigma^2}$$

provides a measure of the proportion of total variation that is explained by between group variability. It is

- 0 when there is no between group variability, $\tau^2 = 0$
- 1 when there is no within group variability $\sigma^2 = 0$

Parameter Estimation: Classical estimation

- Find the MLE's of μ, σ^2, τ^2 from the marginal model for Y_{ij}.
- Use Residual Maximum Likelihood (REML): fit fixed effects by least squares, then estimate variance components by ML using residuals.
- To fit linear mixed effect models in R, use library nlme and function lme

Parameter Estimation: Classical estimation

```
fit.lme = lme(fixed = MathAch ~ 1, random = ~ 1 |School)
summary(fit.lme)

Linear mixed-effects model fit by REML
Random effects:
    Formula: ~1 | School
    (Intercept) Residual
StdDev: 2.934966 6.256862

Fixed effects: MathAch ~ 1
    Value Std.Error DF t-value p-value
(Intercept) 12.63697 0.2443936 7025 51.70747 0

Number of Observations: 7185
Number of Groups: 160
```

Parameter Estimation: Classical estimation

REML Estimates

- $\hat{\mu} = 12.64$
- $\hat{\tau} = 2.93$ or $\hat{\tau}^2 = 8.61$
- $\hat{\sigma} = 6.26$ or $\hat{\sigma}^2 = 39.14$
- $\rho = 8.61/(8.61 + 39.14) = 0.18$

Roughly 20% of the variation in math achievement scores can be attributed to differences among schools. The remaining variation is due to variation among students within schools.

Parameter Estimation: Classical estimation

From a frequentist perspective estimation of the group-specific effects (random effects) is not immediate (unlike within the Bayesian framework).
In *R* we can use the function random.effects () to get estimates

```
random.effects(fit.lme)
          (Intercept)
8367   -6.10301150
8854   -7.35291362
4458   -6.23521734
5762   -7.40281890
6990   -6.13417909
```

Bayesian Hierarchical Model

- Unknown parameters of interest: $\theta_j, \mu, \sigma^2, \tau^2$
- Distribution for θ_j is given by the 2nd level model specification

$$\begin{aligned} Y_{ij} &\sim \mathcal{N}\left(\theta_j, \sigma^2\right) \\ \theta_j &\sim \mathcal{N}\left(\mu, \tau^2\right) \\ \left(\mu, \sigma^2, \tau^2\right) &\sim p(.) \end{aligned}$$

- We specify priors for $\left(\mu, \sigma^2, \tau^2\right)$.
- Can use a default prior $p\left(\mu, \sigma^2, \tau^2\right) \propto 1/\sigma^2$ [a reference prior on τ^2 would cause the posterior to be improper. Use $p\left(\tau^2\right) \propto 1$]
- The joint posterior distribution is given by

$$p\left(\theta_1, \ldots, \theta_m, \mu, \sigma^2, \tau^2 \mid Y\right) \propto p\left(Y \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \mu, \tau^2\right) p\left(\mu, \sigma^2, \tau^2\right)$$

Bayesian Hierarchical Model

replacing variance components with precisions ϕ and ϕ_μ

$$p(Y \mid \theta_j, \phi) \propto \sum_j \sum_i \phi^{1/2} \exp\left(\frac{1}{2}\phi (Y_{ij} - \theta_j)^2\right)$$

$$p(\theta_j, \mid \mu, \phi_\mu) \propto \phi_\mu^{1/2} \exp\left(\frac{1}{2}\phi_\mu (\theta_j - \mu)^2\right)$$

and, under the default prior,

$$p(\mu \mid \theta_j, \phi, \phi_\mu, Y) \propto \prod_j p(\theta_j \mid \mu, \phi_\mu)\, p(\mu) = N\left(\frac{\sum \theta_j}{m}, \frac{1}{m\phi_\mu}\right)$$

See Hoff Chapter 8.3 for full conditionals on $\theta_j, \tau^2, \sigma^2$ (also for the more general case of a Normal-Ga-Ga prior)

Bayesian Hierarchical Model

- We cannot obtain the posterior distributions in closed form.
- We can use Gibbs sampling and create a Markov chain that generates values from the following full conditional distributions:

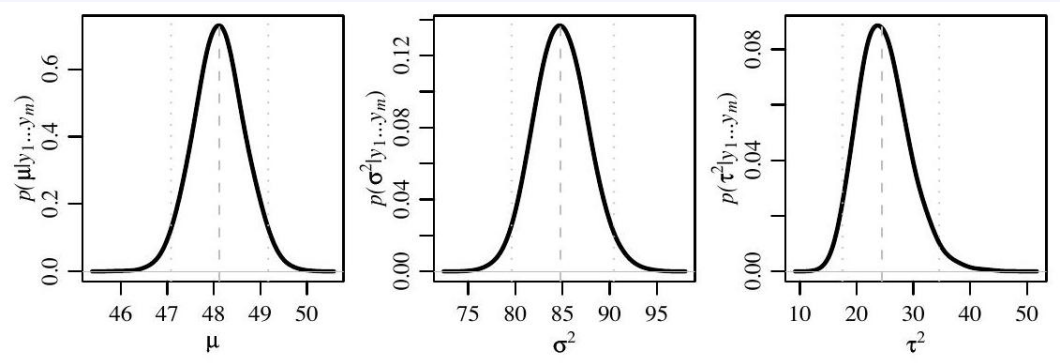
$$p\left(\theta_j \mid Y, \boldsymbol{\theta}_{(-j)}, \mu, \sigma^2, \tau^2\right) \text{ for } j = 1, \ldots, m$$

$$p\left(\mu \mid Y, \boldsymbol{\theta}, \sigma^2, \tau^2\right)$$

$$p\left(\sigma^2 \mid Y, \boldsymbol{\theta}, \mu, \tau^2\right)$$

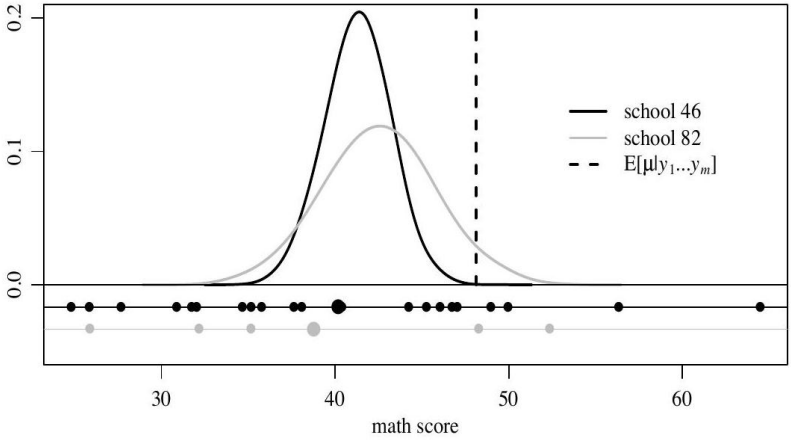
$$p\left(\tau^2 \mid Y, \boldsymbol{\theta}, \mu, \sigma^2\right)$$

Bayesian Hierarchical Model



Posterior means of μ, σ^2, τ^2 are 48.12, 9.21^2 and 4.97^2. Roughly 95% of the scores within a school are within $4 \times 9.21 = 37$ points of each other, roughly 95% of the average school scores are within $4 \times 4.97 = 20$ points of each other.

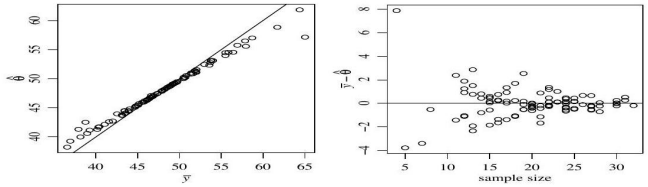
Estimates of group-specific effects



Shrinkage Effect

$$E\left[\theta_j \mid \mu, \tau, Y\right] = \frac{n_j \bar{y}_j / \sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}$$

- average of empirical mean $\bar{y}_j$ and overall mean μ. Pulled away from $\bar{y}_j$ versus μ by amount that depends on n_j
- the larger n_j the more info we have for that group, the less we need to borrow from the rest



Model extension: Unequal Variances

One-way analysis of variance model with unequal variances per group:

$$Y_{ij} = \theta_j + \epsilon_{ij}, \quad \epsilon_{ij} \text{ i.i.d. } N\left(0, \sigma_j^2\right)$$

If we allow each group to have its own mean, why not have different variances? Relax assumption that $\sigma_j = \sigma$ for every j. (θ_j assumed same)

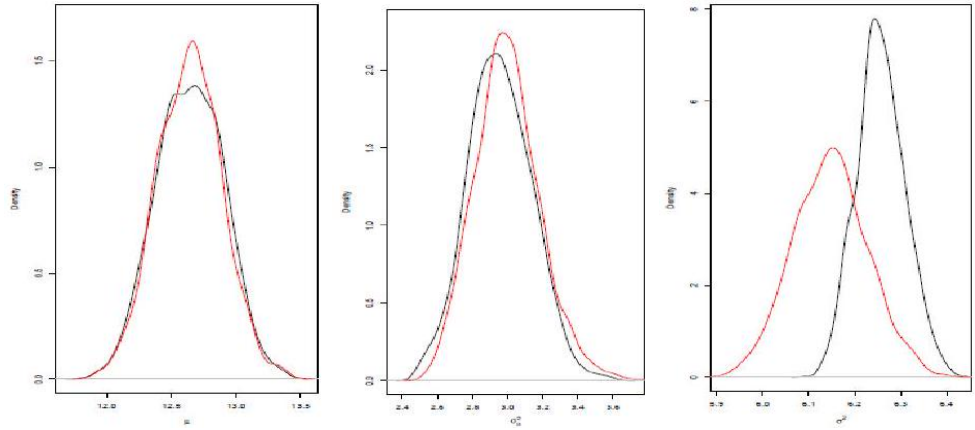
Priors

- $\frac{1}{\sigma_j^2} \sim \text{Gamma}\left(\nu_0/2, \nu_0\sigma_0^2/2\right)$
- prior on σ_0^2 : $p\left(\sigma_0^2\right) \propto \frac{1}{\sigma_0^2}$
- prior on ν_0 : $\nu_0 \sim \text{Gamma}(2, 1/2)$, so prior degrees of freedom 4
- consider $\lambda_j = \frac{\sigma_0^2}{\sigma_j^2}$, so $\lambda_j \sim \text{Gamma}\left(\nu_0/2, \nu_0/2\right)$
- this is equivalent to having the $y_{ij} \mid \theta_j, \sigma_0^2, \lambda_j \sim N\left(\theta_j, \sigma_0^2/\lambda_j\right)$
- marginally the y_{ij} are multivariate t with ν_0 degrees of freedom with means θ_j and scale σ_j^2

Full Conditionals

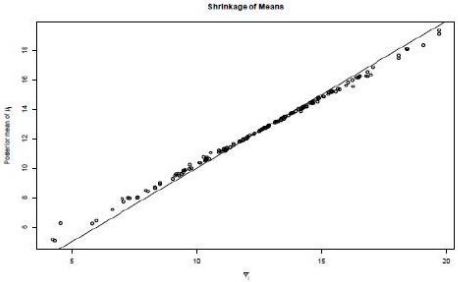
- $\frac{1}{\sigma_j^2} \Big|$ rest $\sim$ Gamma $\left(\left(\nu_0 + n_j \right) / 2, \left(\nu_0 \sigma_0^2 + \sum \left(y_{ij} - \theta_j \right)^2 \right) \right)$
- $\sigma_0^2 \mid$ rest $\sim$ Gamma $\left(\left(a + J\nu_0 \right) / 2, \left(b + \sum \left(1/\sigma_j^2 \right) \right) \right)$
- ν_0 no closed form full conditional

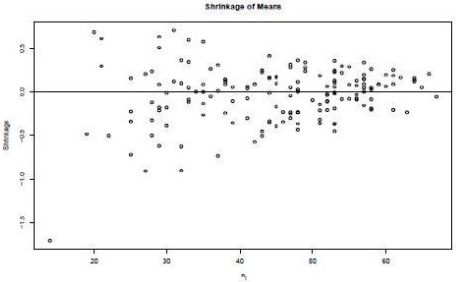
Posterior distributions of μ, τ^2



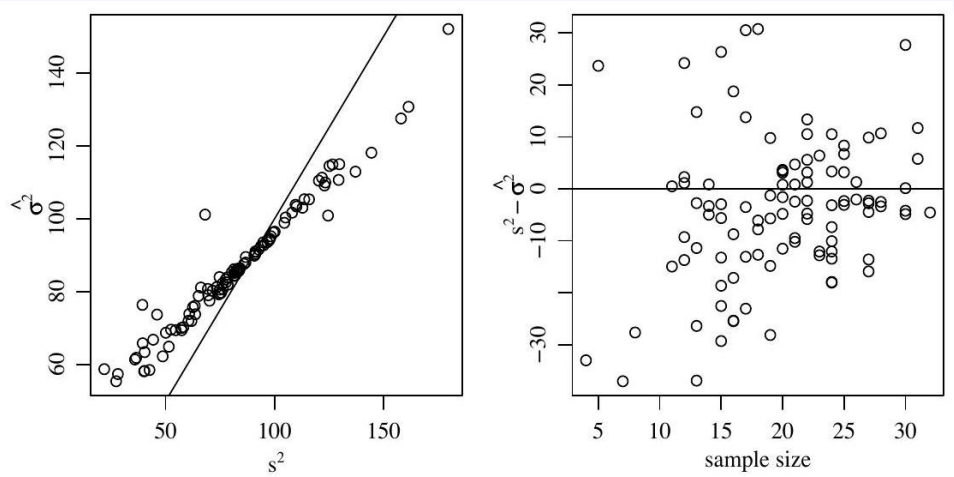
Shrinkage of means

$$E\left[\theta_j \mid \mu, \tau^2, \sigma_0^2, Y\right] = \left(n_j/\sigma_j^2 + 1/\tau^2\right)^{-1} \left(\frac{n_j}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu\right)$$





Shrinkage of variances



Ranking schools

At each iteration, we may rank the schools based on the sampled θ_j and obtain a posterior distribution of the rank of each school.

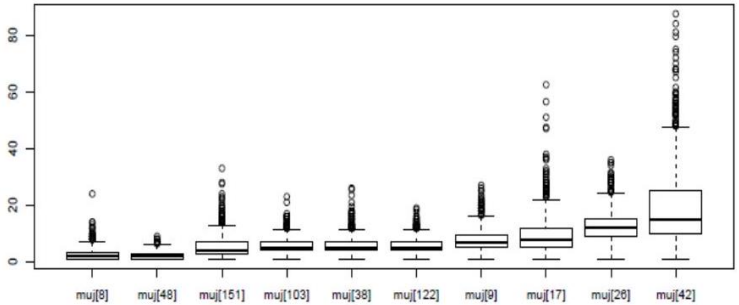


Figure: Figure: Top 10 schools based on rank of sample means

Mixed effect model

- We can write $\theta_j = \mu + \delta_j$ where each school mean is centered at the overall mean μ plus some normal random effect δ_j.
- Substituting this into the distribution for Y_{ij}, we get

$$Y_{ij} = \mu + \delta_j + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$$

with

  - fixed effect, μ
  - school level random effects, δ_j
  - individual random effects ε_{ij}

This leads to a **mixed effects model**.

Mixed Effects Model Parameterization

$$Y_{ij} = \mu + \delta_j + \epsilon_{ij}$$

- The parameters of this model are not identifiable: adding 42 to μ and subtracting 42 from δ_j, leads to a new μ and δ_j, but the same likelihood.
- Model is identifiable with the addition of the prior distributions
- leads to different full conditional distributions
- but this model has very poor mixing!

Summary

- Hierarchical models allow information to be shared across groups of observations.
- We can obtain a posterior distribution for each group mean.
- Each posterior mean will be a convex combination between the observed group mean and the overall mean.
- The group level means are "shrunken" towards the overall mean; the degree of shrinkage depends on the variance components.
- Compromise between fixed effects models
  - each school has its own mean
  - common mean ($\theta_1 = \ldots = \theta_m$)
- Avoids multiple testing