Lecture 7: Hierarchical Modeling

Professor Alexander Franks

2023-03-07

Announcements

- Fill out ESCI evaluations online!
    - What did you like? Would you like to see more courses like this?
    - What can be improved?
- Reading: Chapters 15 and 16

Hierarchical

- Hw 5 out, due 19th.
- Quiz out, due 8th at 2pm. (MCMC)
  (1 quiz next).

Comparing Multiple Related Groups

- Hierarchy of nested populations
- Models which account for this are called *hiearchical* or *multi-level* models

Some examples:

- Patient outcomes within several different hospitals
- People within counties in the United States (e.g. Asthma mortality example)
- Athlete performance in sports
- Genes within a group of animals

Eight schools example

- A study was performed for the Educational Testing Service (ETS) to evaluate the effects of coaching programs on SAT preparation
- Each of eight different schools used a short-term SAT prep coaching program
- Compute the average SAT score in those who did take the program minus those that did not participate in the program
- We observe the average difference varies by school. What accounts for these differences?

50 students per school, randomly assign 25

Y_i, $i = 1 \ldots 8$

difference in SAT from training vs control.

- Access varies by school
- Income
- Quality of the school
- populations are different
- Motivation
- Language.

- Chance!
  + By Chance strong students in 1 group & weaker in the other.

Eight schools example

- Interested in "real" differences due to training
- Want to reduce ~~effect~~ influence of chance variability
- How do we estimate the effect of the program in each of the schools?
- Two extremes:
  - Estimate the effect of the program in every school independently
    - A separate prior distribution for each school effect
  - Or assume the effect is the same in every school
    - Combine all the data
  - A compromise between the above 2 options?

Eight Schools Example

$$y_j \sim N(\theta_j, \sigma_j^2)$$, $\theta_j \sim N(\mu, \frac{\sigma^2}{K})$

- y_j is the observed effects of the program in school j
  - Based on a sample of test scores from those in the program and those not in the program
- θ_j are the true *unknown* effects of the program in school j
- Variances, σ_j^2, are *known*
  - Determined by the number of students in the sample

$\sigma_j^2 = \frac{\sigma^2}{n_j}$

$P(\theta_j \mid y_j) \sim N(w y_j + (1-w)\mu, w\sigma_j^2)$

$w = \frac{n_j}{n_j + K}$

Eight Schools Example

```
J <- 8
y = c(28,  8, -3,  7, -1,  1, 18, 12)
sigma <- c(15, 10, 16, 11,  9, 11, 10, 18)
```

σ_j's

- Assuming the effect of the program on each school is identical.
- What are the chances of seeing a value as large as 28?
- As small as -3?

$$y_j \sim N(\theta, \sigma_j^2)$$

Why is MLE not $\bar{y}$?

$$L(\theta) \propto \prod_{i=1}^{8} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{\frac{-(y_j-\theta)^2}{2\sigma_j^2}}$$

$$\hat{\theta}_{MLE} = \sum_{i=1}^{8} 1/\sigma_j^2 \, y_j \Big/ \sum 1/\sigma_j^2$$

Eight Schools Example

If effects are actual equal, what is it?

```
## Compute the precision frome each school
prec <- 1/sigma^2

## global estimate is a weighted average
mu_global <- sum(prec * y / sum(prec))
mu_global
```

```
## [1] 7.685617
```

$\hat{\theta}_{MLE} \approx 7.7$ if no variation between schools.

Eight Schools Example

- Assume the effect of the program on each school is identical, i.e. $\theta_j = \theta$ $= 7.7$
- What are the chances of seeing a value as large as 28?
- As small as -3?

```
# 1000 datasets with mean mu_global but different sigmas

## Chance of seeing a value greater than 28
mean(sapply(1:1000, function(x)
  max(rnorm(J, mean=mu_global, sd=sigma))) >= 28)
```

```
## [1] 0.358
```

$= 7.7$ $\leftarrow \sigma_j$

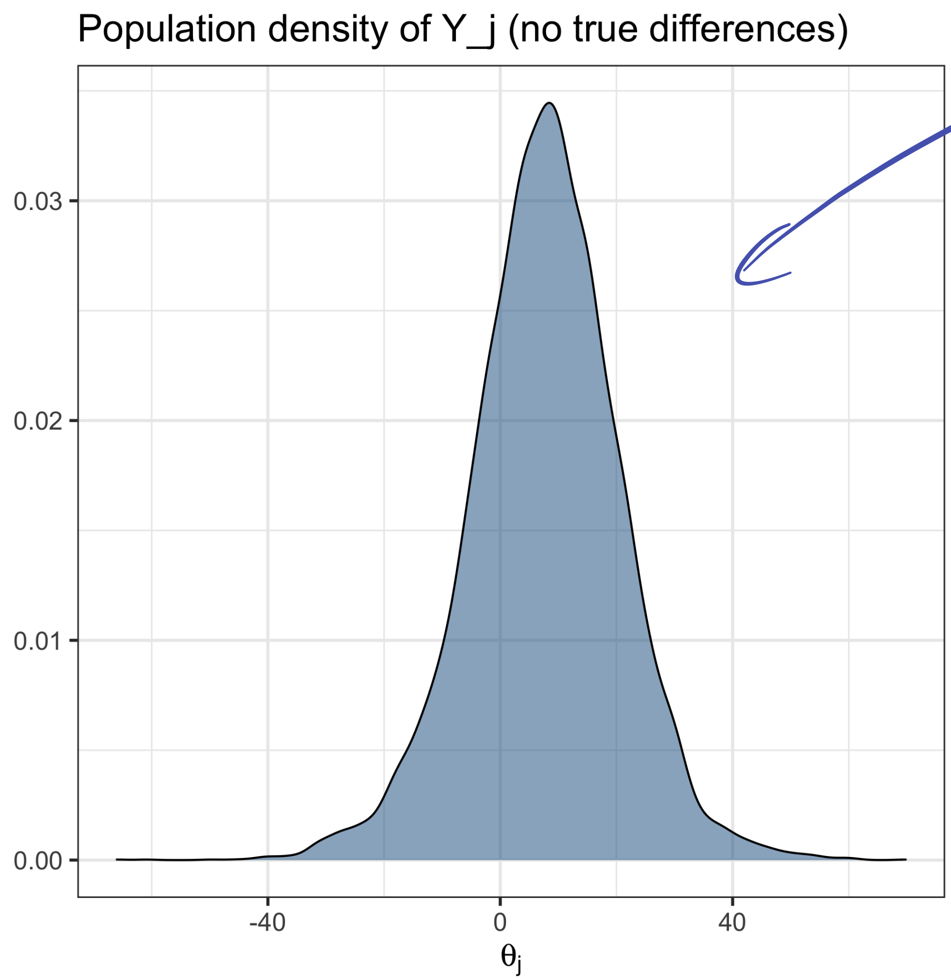
```
## Chance of seeing a value less than -3
mean(sapply(1:1000, function(x)
  min(rnorm(J, mean=mu_global, sd=sigma))) <= -3)
```

```
## [1] 0.799
```

Eight Schools Example

7.7

Density of $Y_j \sim N(\theta_{\text{pooled}}, \sigma_j^2)$



How much y can vary due to chance.

Eight Schools Example

$$y_j \sim N(\theta_j, \sigma_j^2)$$

- θ_j are the true unknown effects of the program in school j
- y_j is the observed effects of the program in school j
    - Based on a sample of test scores from those in the program and those not in the program
    - Number of people in the sample determine the magnitude of σ_j^2

$$\hat{\theta}_{j,MLE} = y_j$$

Eight Schools Example

- How do we estimate θ_j?
  - Independent: $\hat{\theta}_j^{(MLE)} = y_j$ is the MLE
  - Identical effects: $\hat{\theta}_j^{(pool)} = \frac{\sum_i \frac{1}{\sigma_i^2} y_i}{\sum \frac{1}{\sigma_i^2}} = \theta^{(pool)}$ for all j ≈ 7.7.
    - Same effect for all schools: estimate using a weighted average of the observed effects

Eight Schools

Bias - Variance Tradeoff.

```
theta_j_mle <- y
theta_j_mle
```

```
## [1] 28  8 -3  7 -1  1 18 12
```

```
theta_j_pooled <- rep(sum(1/sigma^2 * y) / sum(1/sigma^2), J)
theta_j_pooled
```

```
## [1] 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617 7.685
```

Is there a middle ground between two extremes?

$$\hat{\theta}_{j,\,partial} = \tilde{w} y_j + (1-\tilde{w})\,\theta_{pooled}$$

Eight schools example

A hierarchical model:

How variable are the true SAT effects across schools.

θ_{pooled}

$$\boxed{\begin{aligned}\theta_i &\sim N(\mu, \tau^2) \\ y_j &\sim N(\theta_j, \sigma_j^2)\end{aligned}}$$

$\theta_{j,pm} = w y_j + (1-w)\mu$

$w = \dfrac{1/\sigma_j^2}{1/\sigma_j^2 + \frac{1}{\tau^2}}$

random (not known)

- Add a *shared* normal prior distribution to θ_j
- Assume the global mean, μ is also unknown parameter.
- How do we choose prior for μ?

$\mu \sim N(\mu_0, \tau_0^2)$? or $p(\mu) \propto 1$? (hyperprior)

- Need to estimate all of $(\mu, \theta_1, \ldots, \theta_8)$ with MCMC

9-parameter posterior.

- τ^2 determines how much weight weight we put on the independent estimate vs the pooled estimate

$P(\mu, \theta_1, \ldots \theta_8 \mid y_1, \ldots y_8, \sigma_j)$

$$P(\mu, \theta_1, \ldots \theta_8 \mid y_1, \ldots y_8, \sigma_j\text{'s}) \propto$$

$$\underbrace{\prod_{i=1}^{8} \left[\frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(y_j-\theta_j)^2}{2\sigma_j^2}} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta_j-\mu)^2}{2\tau^2}} \right] P(\mu)}$$

- Compromise: $\hat{\theta}_j^{(shrink)} = w\theta_j^{(MLE)} + (1-w)\theta_j^{(pool)}$
- How is this different from the standard normal-normal model we saw before?

Intuition behind shrinkage

- $Y_j = \theta_j + \epsilon_j$ and for simplicity assume that the variance of ϵ, σ^2 for all j
  - θ_j represents true differences between schools (signal)
  - ϵ_j is sampling variability (noise, chance variation)
- $Var(Y_j) = Var(\theta_j) + Var(\epsilon_j) = \tau^2 + \sigma_j^2$
  - The variance of the observed outcomes is the sum of signal variance, τ^2, and the sampling variance σ_j^2
- Consequence: the observed outcomes always have higher variance across groups than the signal
  - $\mathrm{Var}(Y_j) > \mathrm{Var}(\theta_j)$
- Intuition: reduce the variance by shrinking Y_j's closer together!
  - Want the variance of the shrunken estimates to be close to τ^2

Eight Schools examples

$\theta_j \sim N(\mu, \tau^2)$

Comments:

- The global average, μ, is a parameter so also has uncertainty
- How dow we determine how much to shrink, e.g. how do we determine τ^2? → signal variability.
- Is the training program effective in school j?
  - What is $P(\theta_j > 0 \mid y, \sigma_j)$?
- On aveage (over all schools) is the training program effective?
  - What is $P(\mu > 0 \mid y)$?

Eight schools example

- If τ^2 is large, the prior for θ_j is not very strong
  - If $\tau^2 \to \infty$ equivalent to the no pooling model
- If τ^2 is small, we assume a priori that θ_j are very close
  - if $\tau^2 \to 0$ equivalent to the complete pooling model, $\theta_j = \mu$

$$\theta_{j,pm} = w y_j + (1-w)\mu$$

$$w = \frac{1/\sigma_j^2}{1/\sigma_j^2 + \frac{1}{\tau^2}}$$

Estimating parameters

- MH: Need to generate a proposal from a 9-dimensional posterior distribution
  - Eight parameters for θ_j and one for μ
- Gibbs: sample each of the 9 parameters from the complete conditionals
  - Sample $p(\theta_j \mid \theta_{-j}, \mu)$
  - Sample $p(\mu \mid \theta_1, \ldots \theta_8, \mu)$
- Stan

Eight Schools Estimation

```
J <- 8
y = c(28,  8, -3,  7, -1,  1, 18, 12)
sigma <- c(15, 10, 16, 11,  9, 11, 10, 18)

## fix the variance of the prior to a number
tau <- 20
```

Eight Schools in Stan

```
// saved as 8schools.stan
data {
  int<lower=0> J;          // number of schools
  real y[J];               // estimated treatment effects
  real<lower=0> sigma[J]; // standard error of effect estimates
  real<lower=0> tau;       // shrinkage standard deviation
}
parameters {
  real mu;            // population treatment effect
  vector[J] eta;   // unscaled deviation from mu by school
}
transformed parameters {
  vector[J] theta = mu + tau * eta;   // school treatment effects
}
model {
  target += normal_lpdf(eta | 0, 1);          // prior log-density
  target += normal_lpdf(y | theta, sigma); // log-likelihood
}
```

A number I choose.

(global average)

theta

θ_j

y ~ normal(theta, sigma)

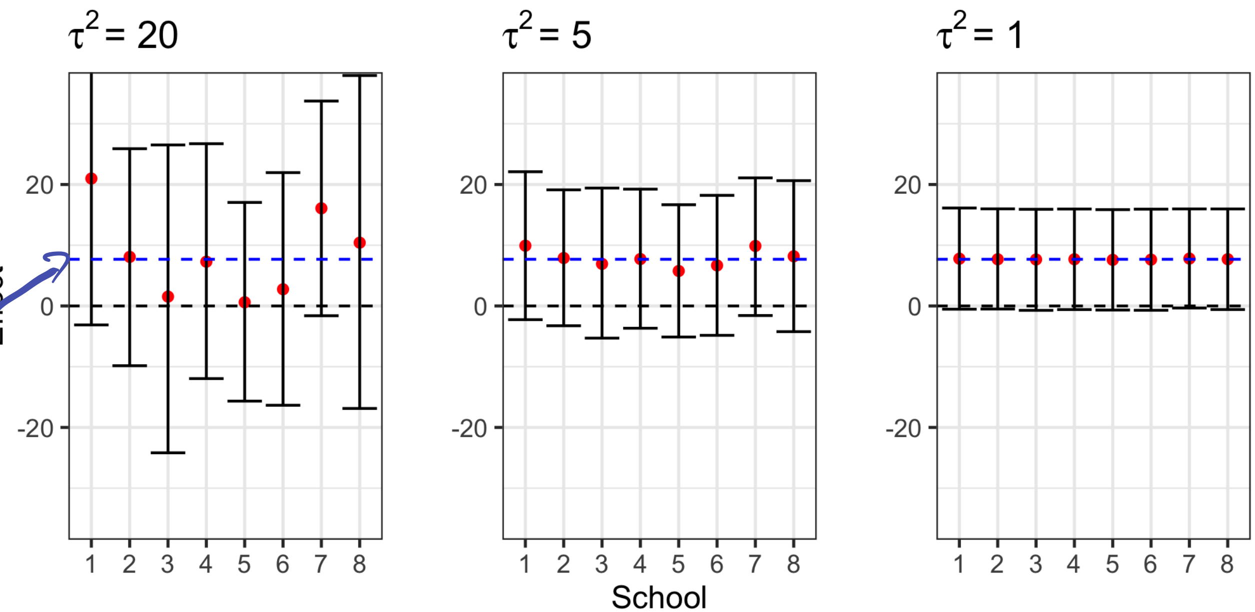
theta ~ normal(mu, tau)

(mu $\propto$ const, flat a prior)

Eight Schools example

θ_j

μ



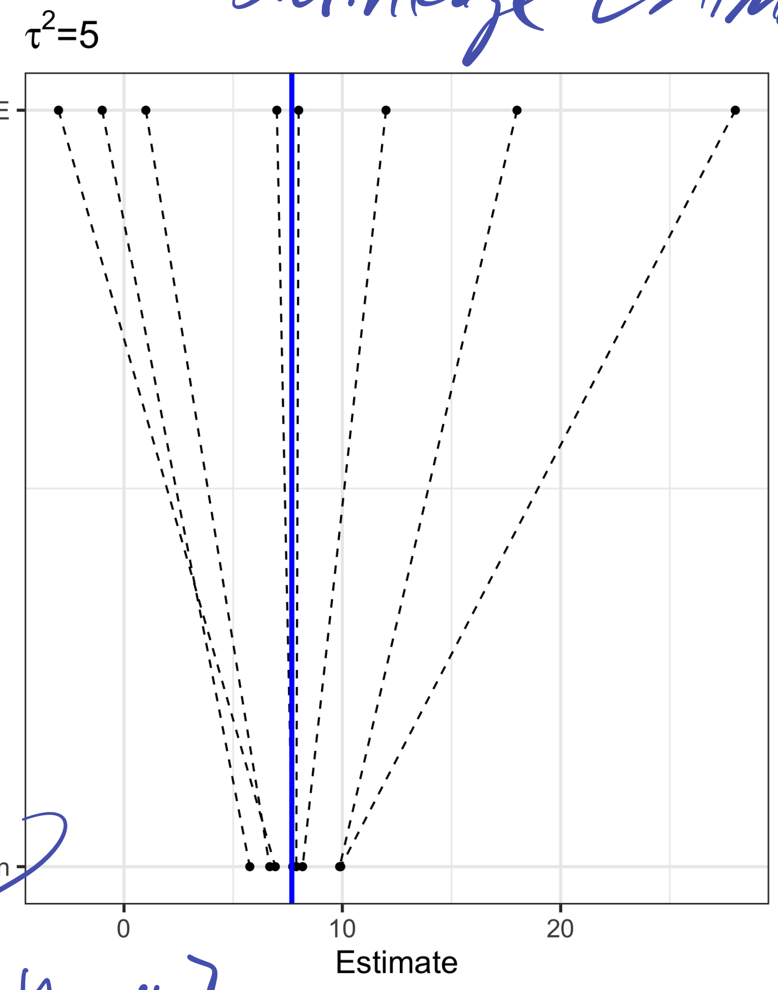
- Larger τ^2 means more variability in θ_j.
- Larger τ^2 means more posterior uncertainty.

$$\theta_j \sim N(w y_j + (1-w)\mu, \, w\sigma_j^2)$$

MLE vs Posterior Mean

Shrinkage Estimation.

$\hat{\theta}_{MLE}$



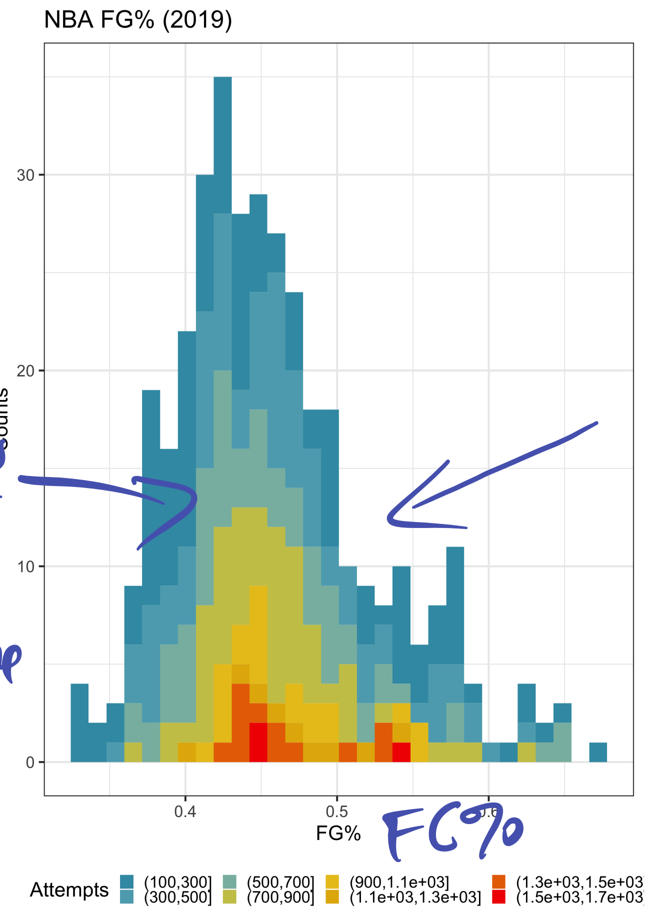
$E[\theta_i \mid y_1, \ldots y_8]$

Basketball Example

θ_j is FG%

y_j ~ Fraction of makes

μ is league average FG%



FG%