Lecture 5: Hierarchical Modeling

Professor Alexander Franks

Announcements

- Reading: Chapter 5 of BDA
- Hwk due Sunday.
- Ch. 10/11

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?

$Y_1, \quad Y_2, \ldots\ldots Y_8$

- Sampling variability.
- Program effectiveness varies by school
  + Demographics
  + School Funds
  + Instructor variation.

Eight schools example

- Interested in “real” differences due to training
- Want to reduce effect of chance variability
- How do we estimate the effect of the program in each of the schools?

Eight schools example

- Consider 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

      More Data.
  - A compromise between the above 2 options?

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

σ_j^2 is known

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

No Pooling Model

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

Eight Schools Example

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

- 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
- Assume variances, σ_j^2, are *known*
  - e.g. determined by the number of students in the sample

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)
```

- 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?

$Y_j \sim N(\mu, \sigma_j^2)$

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

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

No Pooling.

Eight Schools Example

- Assume the effect of the program on each school is identical, i.e. $\theta_j = \mu$
- Assume a flat prior on μ, what is $p(\mu \mid y_1, \ldots, y_8, \sigma_1, \ldots \sigma_8)$?

$p(\mu) \propto \text{const}$

$$p(\mu \mid y_1, \ldots y_8, \sigma_1, \ldots \sigma_8) \propto L(\mu) \propto$$

$$\prod_{i=1}^{8} \cancel{\frac{1}{\sqrt{2\pi\sigma_i^2}}} e^{-\frac{(y_j-\mu)^2}{2\sigma_i^2}} \propto \exp\left[-\sum \frac{1}{2\sigma_i^2}(y_i-\mu)^2\right]$$

$$\propto \exp\left(-\sum^{8} \frac{1}{2\sigma_i^2}\left(\mu^2 - 2y_i\mu\right)\right)$$

Check

$$\propto \exp\left(-\sum \frac{1}{2\sigma_i^2}\left(\mu - \frac{\sum \frac{y_i}{\sigma_i^2}}{\sum 1/\sigma_i^2}\right)^2\right)$$

$$\longrightarrow \mathcal{N}\left(\sum w_i y_i,\ \frac{1}{\sum 1/\sigma_i^2}\right)$$

$$\boxed{w_i = \frac{\frac{1}{\sigma_i^2}}{\sum 1/\sigma_i^2}} \qquad \hat{\mu}_{MLE} = \sum w_i y_i$$

<u>Fisher Weighting</u>

Eight Schools Example

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

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

```
[1] 7.685617
```

Eight Schools Example

- Assume the effect of the program on each school is identical, i.e. $\theta_j = \mu$
- What are the chances of school 1 having an effect large as 28 (given $\sigma_1 = 15$)?
- Y_3 as small as -3 (given $\sigma_3 = 16$)?

$$\int P(\tilde{Y} \geq 28 \mid \mu, \sigma = 15) P(\mu \mid y_1, \dots y_8, \dots) d\mu$$

$$= P(\tilde{Y} \geq 28 \mid \sigma = 15, y_1, \dots y_8)$$

$$P(\tilde{Y} < -3 \quad \sigma = 6 \quad , \dots \quad 8$$

$P(\mu|y)\ \mathcal{N}\left(\sum w_i y_i,\ \frac{1}{\sum 1/\sigma_i^2}\right)$

$P(\tilde{y}|\mu) \sim \mathcal{N}(\mu, \sigma^2)$

$$y = \mu + \varepsilon_1 \longrightarrow y = \sum w_i y_i + \varepsilon_1 + \varepsilon_2$$

$$\mu = \sum w_i y_i + \varepsilon_2$$

$$y \sim \mathcal{N}\left(\sum w_i y_i,\ \frac{1}{\sum 1/\sigma_i^2} + \sigma^2\right)$$

Posterior Prediction Under Complete Pooling

```
prec <- 1/sigma^2

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

print(sprintf("mu is %f", mu_global))
```

```
[1] "mu is 7.685617"
```

```
1 - pnorm(28, mean=mu_global, sd=sqrt(1/sum(1/sigma^2)+sigma[1]^2))
```

```
[1] 0.09560784
```

```
pnorm(-3, mean=mu_global, sd=sqrt(1/sum(1/sigma^2)+sigma[3]^2))
```

```
[1] 0.2587447
```

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

Eight Schools Example

How do we estimate θ_j?

- Assume effects are totally independent: $\hat{\theta}_j^{(MLE)} = y_j$ is the MLE
- Assume effects are identical: $\hat{\theta}_j^{(pool)} = \frac{\sum_i \frac{1}{\sigma_i^2} y_i}{\sum \frac{1}{\sigma_i^2}}$
  - Same effect for all schools: estimate using a weighted average of the observed effects

Eight Schools

```
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 7.685617
```

Partial Pooling

- Compromise: $\hat{\theta}_j^{\text{shrink}} = w\theta_j^{\text{MLE}} + (1 - w)\theta^{pooled}$

$0 \leq w \leq 1$

Eight schools example

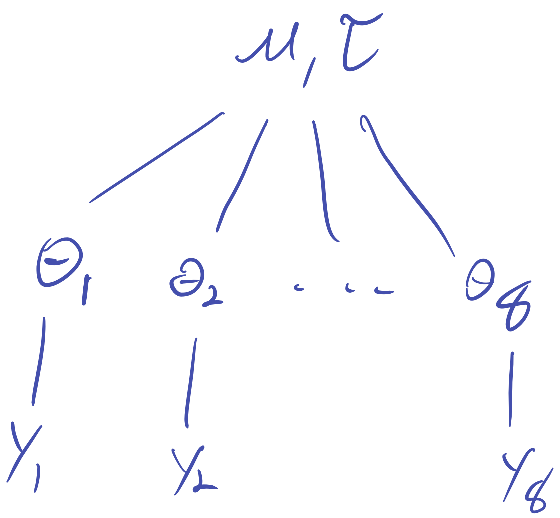
Add a *shared* normal prior distribution to θ_j

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

- The global mean, μ, is also an unknown parameter. What prior should we choose?
- τ^2 determines how much weight weight we put on the independent estimate vs the pooled estimate.
- A 9-parameter posterior: $p(\mu, \theta_1, \ldots, \theta_8 \mid y_1, \ldots, y_8, \sigma_1, \ldots, \sigma_8)$

If μ and τ^2 known

What is $P(\theta_1 \mid y_1, \cancel{\ldots, y_8}, \underline{\mu, \tau^2})$?



$$\sim N\left(wY_1 + (1-w)\mu, \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\tau^2}}\right)$$

$$w = \frac{1/\sigma_1^2}{1/\sigma_1^2 + \frac{1}{\tau^2}}$$

Intuition for shrinkage

- $Y_j = \theta_j + \epsilon_j$
  - For simplicity assume $Var(\epsilon_j) = \sigma^2$ for all j
  - θ_j represents true effect in school j (signal)
  - $Var(\theta_j) = \tau^2$ represents how much the true effects vary across schools
    - ϵ_j is sampling variability (noise, chance variation)
- $\hat{\theta}^{\text{unpooled.}}_{MLE} = Y_j$

28, 10, -3, 5, ...

$\hat{\theta} = Y_i = \theta_i + \epsilon_i$

$Var(\hat{\theta}) \simeq Var(\theta_i) + Var(\epsilon_i) = \tau^2 + \sigma^2$

Intuition for shrinkage

- Consequence: the observed outcomes always have higher variance than the signal, i.e. $\mathrm{Var}(Y_j) > \mathrm{Var}(\theta_j)$
- Intuition: reduce the variance by shrinking estimates to a common mean!
- The variance of the shrunken estimates should be close to τ^2

Eight schools example

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?
- Is the training program effective in school j?
  - What is $P(\theta_j > 0 \mid y)$?
- On average (over all schools) is the training program effective?
  - What is $P(\mu > 0 \mid y)$?

$P(\tilde{y} > 0 \mid y)$

posterior predictive distn.

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, $\theta_j = \mu$

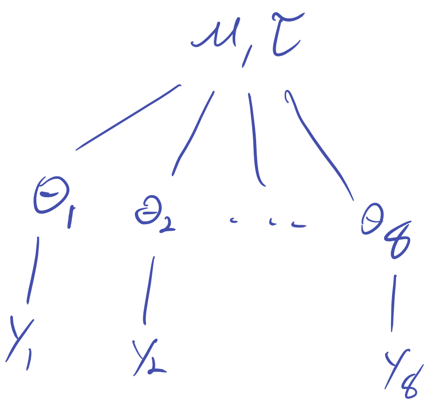
$$w_i = \frac{1/\sigma_i^2}{1/\sigma_i^2 + 1/\tau^2}$$

Inference

- Factorize the density into tractable components
  - $p(\mu \mid y_1, \ldots, y_8, \tau^2)$
  - $p(\theta_i \mid \mu, y_i)$
- Later: MCMC or other approximate methods

$$P(\theta, \mu) = P(\mu) P(\theta \mid \mu)$$

$$P(\mu, \theta_1, \ldots \theta_8 \mid y_1, \ldots y_8) = P(\mu \mid y_1, \ldots y_8) \times \prod^{8} P(\theta_i \mid \mu, y_i)$$

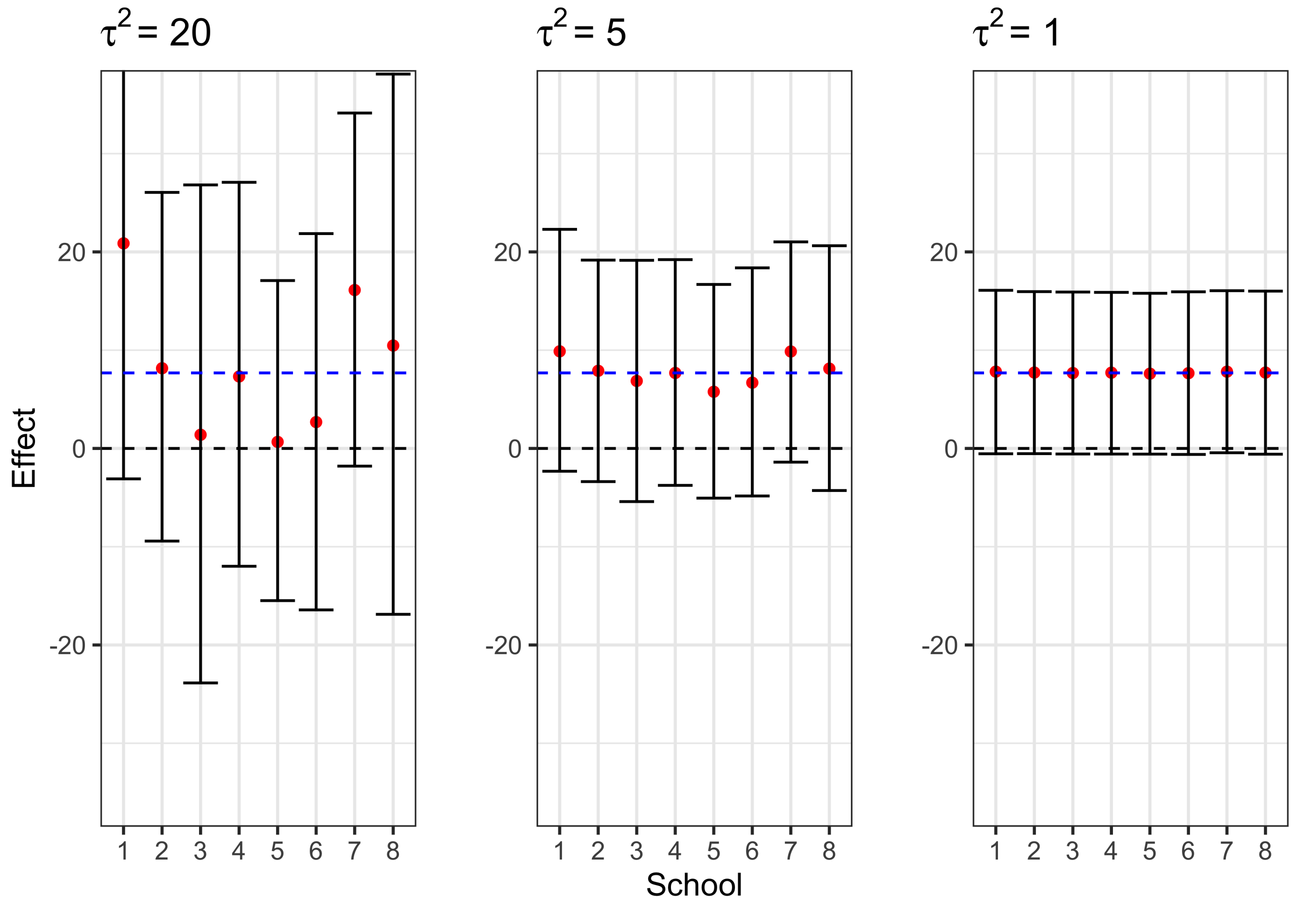


$$P(\mu \mid y_1, \ldots y_8) \rightarrow N\left(\sum w_i y_i, \frac{1}{\sum 1/\sigma_i^2}\right)$$

$$P(\theta_i \mid y_i, \mu) \rightarrow N\left(\underline{1/\sigma_i^2 y_i} + (\,)\mu, \;\cdot\; \right)$$

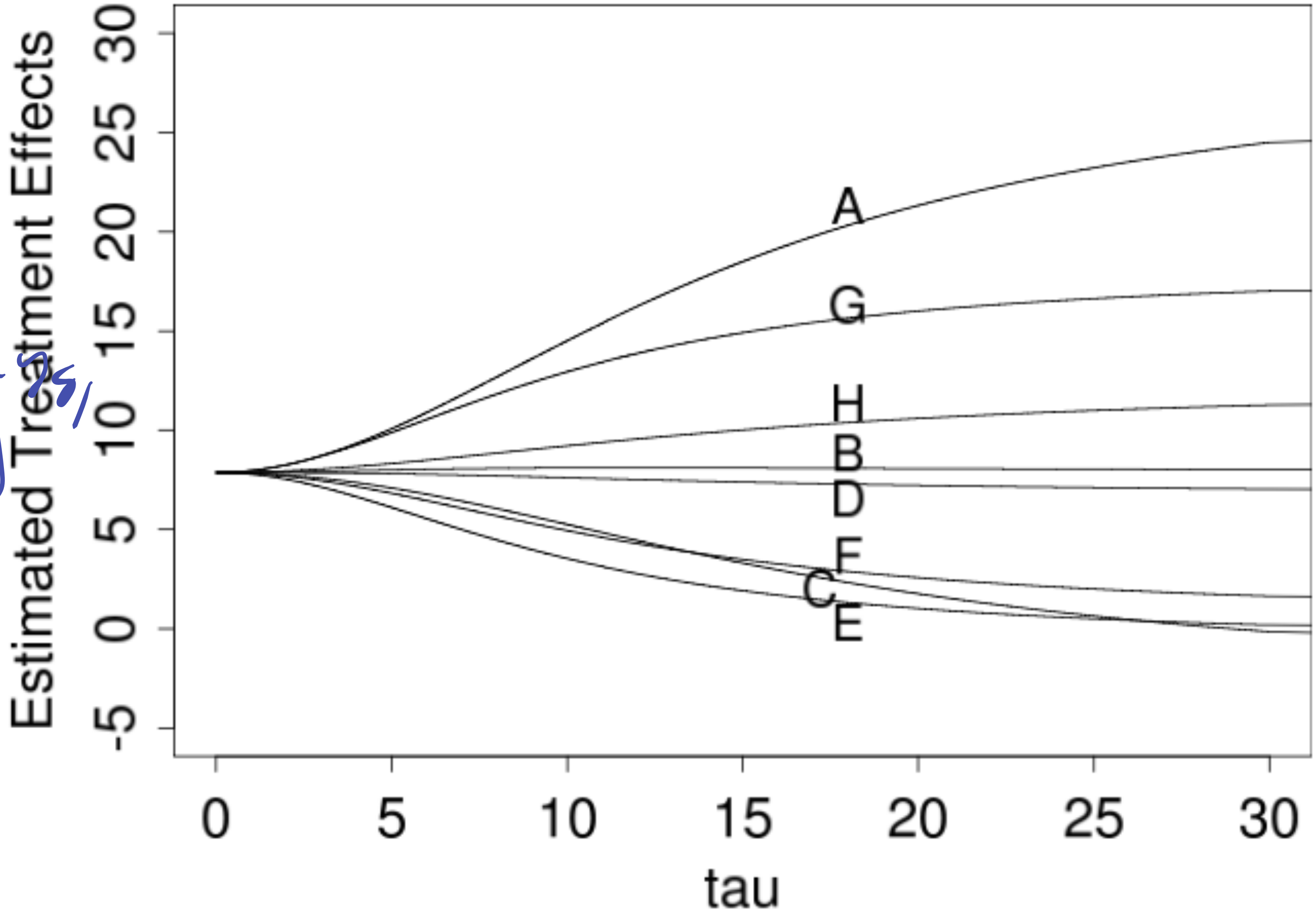
$1/\sigma_i^2 + 1/\tau^2$

Eight Schools example

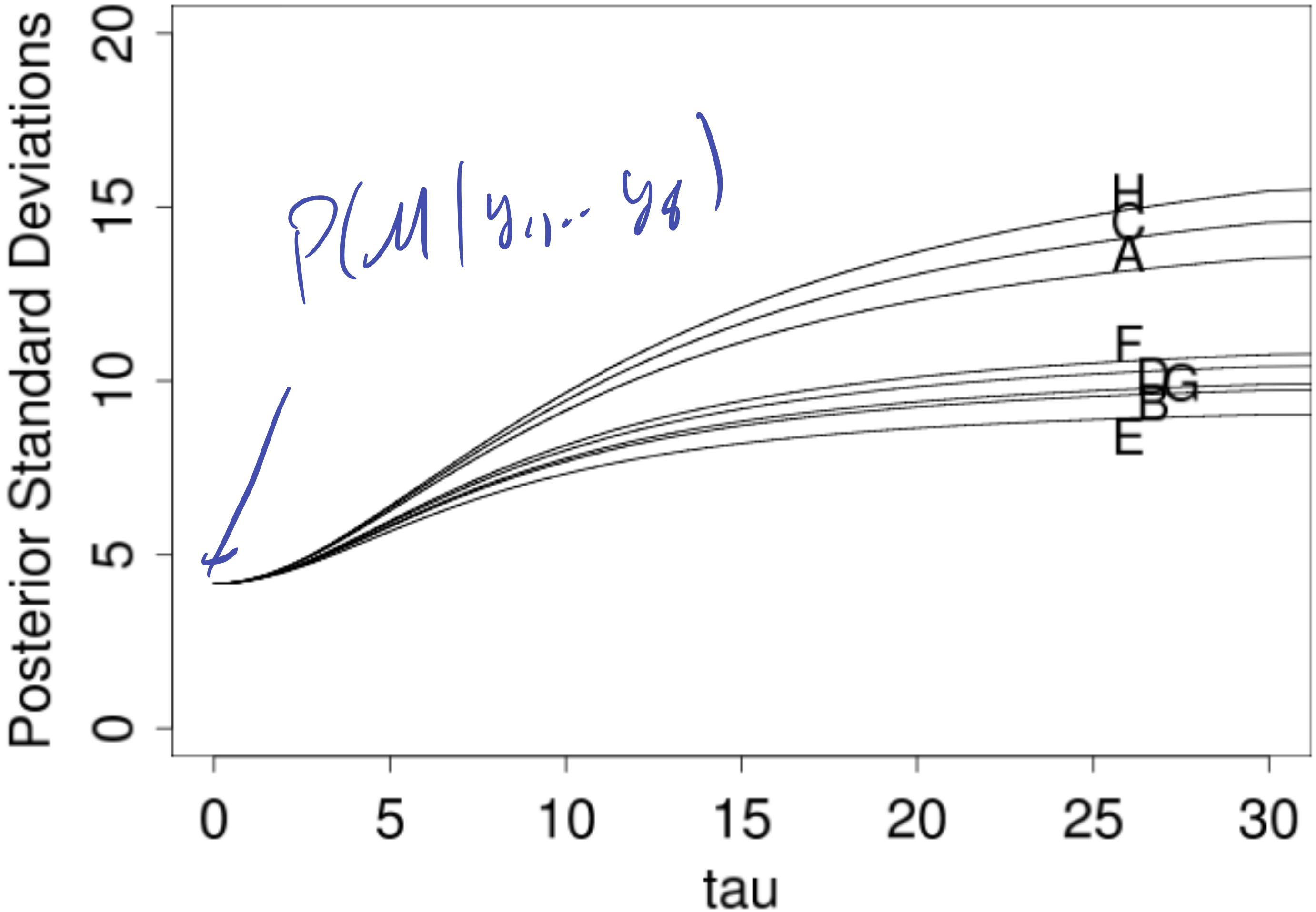


The impact of τ

$E[\theta_i \mid y_1 \ldots y_J, \tau]$



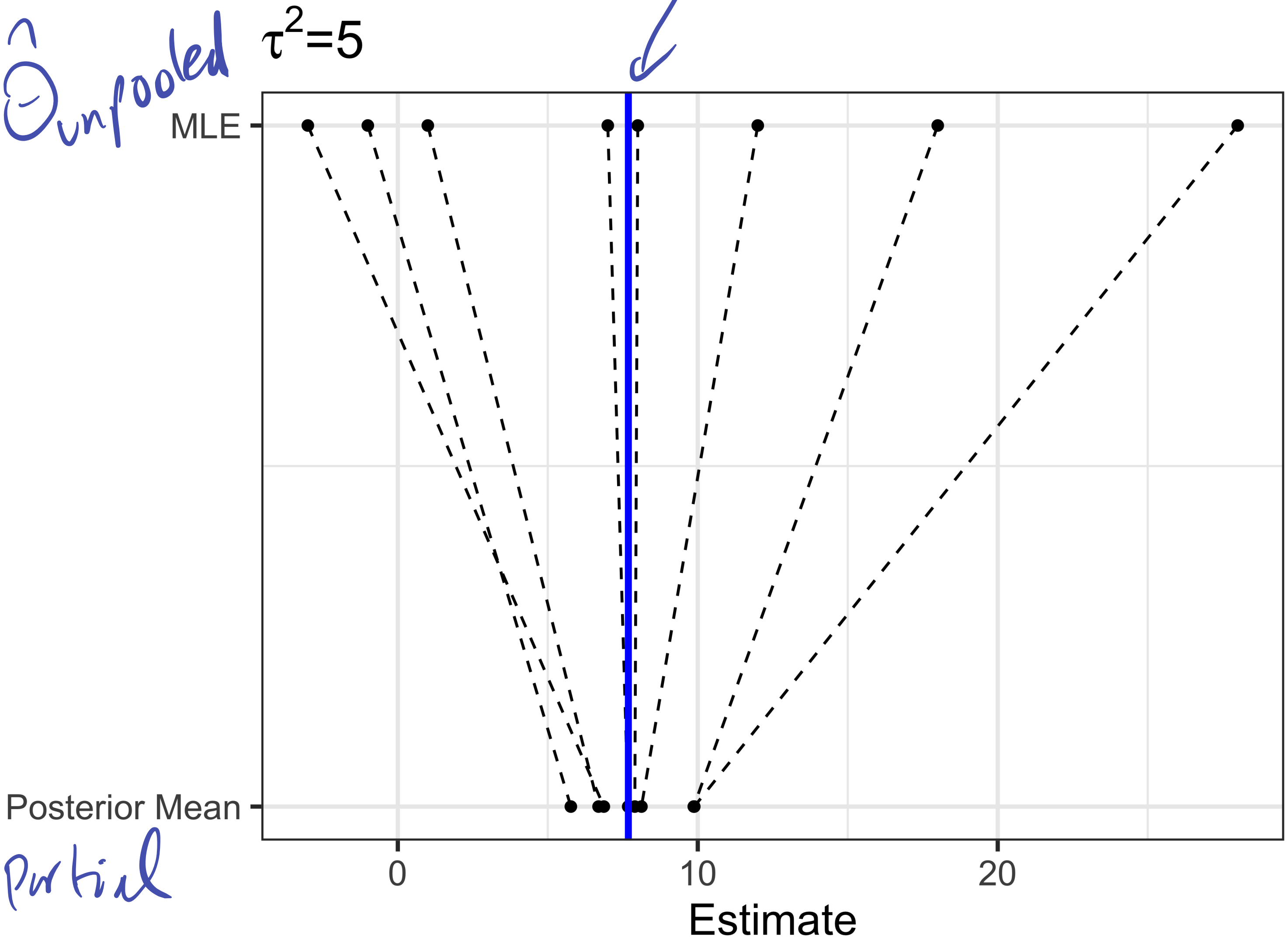
The impact of τ



MLE vs Posterior Mean

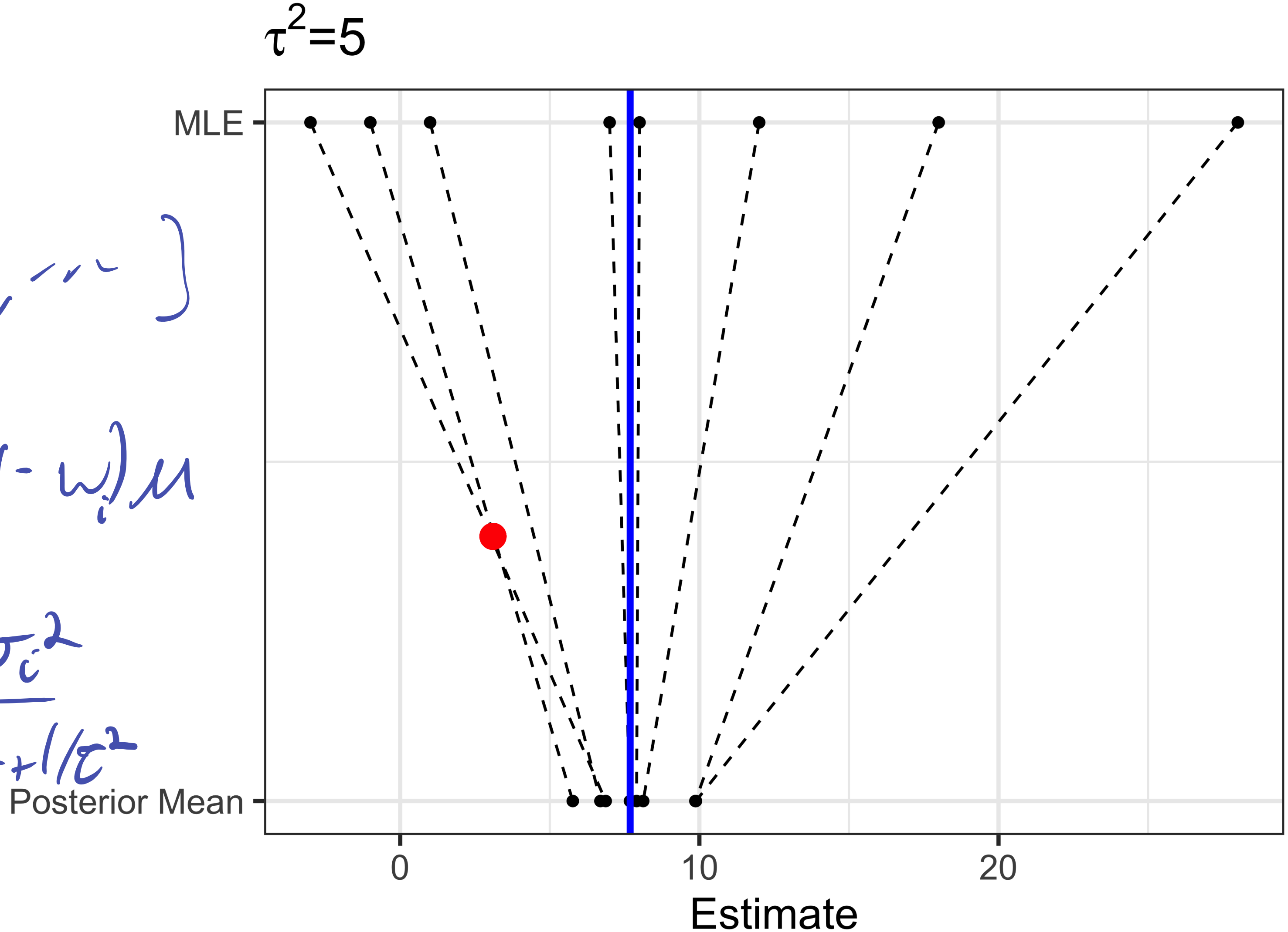
$\hat{\theta}_{complete}$

$\hat{\theta}_{unpooled}$



Partial

MLE vs Posterior Mean



$E[\theta_i \mid \mu, \ldots]$

$w_i Y + (1 - w_i)\mu$

$w_i = \dfrac{1/\sigma_i^2}{1/\sigma_i^2 + 1/\tau^2}$