(5.4,5.5) ETS Experiments and Hierarchical Normal Models

Weihang Ren

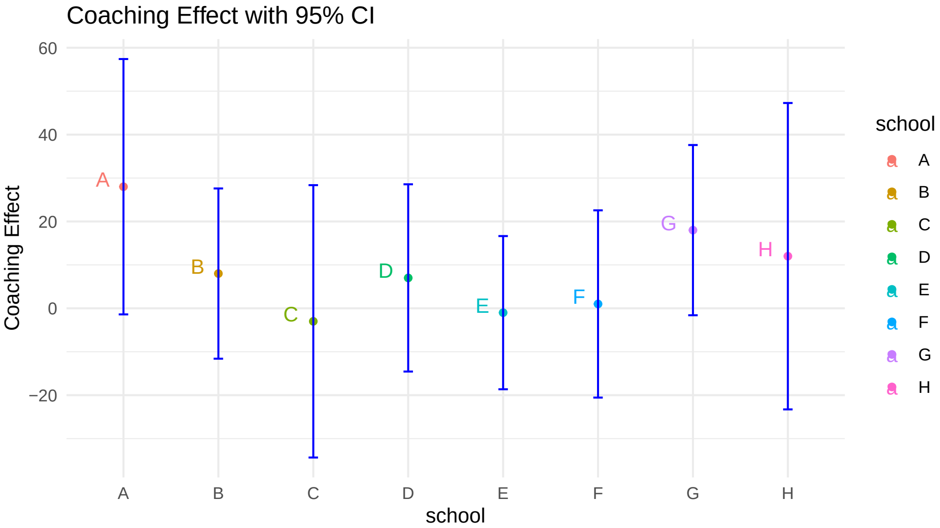
The ETS Dataset

- ETS performed a study to analyze the effects of coaching programs on test scores (SAT Verbal).
- Data was collected from eight high schools.
- Data:
  - The estimated coaching effect $y_j, j = 1, 2, \cdots, 8$,
  - And its sampling variance σ_j^2.
  - Approximated normal sampling distribution.

High School	A	B	C	D	E	F	G	H
Coaching Effect	28	8	-3	7	-1	1	18	12
SE of C.E.	15	10	16	11	9	11	10	18

- Question: How effective are SAT-V prep courses?

The ETS Dataset



The ETS dataset : Data Structure

- J independent experiments (schools).
- Experiment j is to estimate θ_j with n_j data points y_{ij} (Coaching Effect).

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2), i = 1, \cdots, n_j, j = 1, \cdots, J$$

  σ^2 is assumed to be known.

- Denote $\bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$, then

$$\bar{y}_{\cdot j}|\theta_j \sim N(\theta_j, \sigma_j^2), \sigma_j^2 = \sigma^2 / n_j$$

Nonhierarchical Approaches

Separate Estimates

Pooled Estimates

Nonhierarchical Approaches

Separate Estimates

- Consider each estimate θ_j separately.

$$\bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

Pooled Estimates

- Consider a single common effect

$$\bar{y}_{\cdot\cdot} = \frac{1}{\sum_{j=1}^{J} n_j} \sum_{i,j} y_{ij} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \bar{y}_{\cdot j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}$$

Nonhierarchical Approaches

Separate Estimates

- Consider each estimate θ_j separately.

$$\bar{y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

Pooled Estimates

- Consider a single common effect

$$\bar{y}_{\cdot\cdot} = \frac{1}{\sum_{j=1}^{J} n_j}\sum_{i,j} y_{ij} = \frac{\sum_{j=1}^{J}\frac{1}{\sigma_j^2}\bar{y}_{\cdot j}}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2}}$$

	df	SS	MS	$\mathrm{E}(\mathrm{MS}\mid\sigma^2,\tau)$
Between Groups	$J-1$	$\sum_i\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot j})^2$	$SS/(J-1)$	$n\tau^2+\sigma^2$
Within Groups	$J(n-1)$	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot j})^2$	$SS/J(n-1)$	σ^2
Total	$Jn-1$	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot\cdot})^2$	$SS/(Jn-1)$	

Nonhierarchical Approaches

Separate Estimates

- Consider each estimate θ_j separately.

$$\bar{y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

Pooled Estimates

- Consider a single common effect

$$\bar{y}_{\cdot\cdot} = \frac{1}{\sum_{j=1}^{J} n_j}\sum_{i,j} y_{ij} = \frac{\sum_{j=1}^{J}\frac{1}{\sigma_j^2}\bar{y}_{\cdot j}}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2}}$$

	df	SS	MS	$\mathrm{E}(\mathrm{MS}\mid\sigma^2,\tau)$
Between Groups	$J-1$	$\sum_i\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot j})^2$	$SS/(J-1)$	$n\tau^2+\sigma^2$
Within Groups	$J(n-1)$	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot j})^2$	$SS/J(n-1)$	σ^2
Total	$Jn-1$	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot\cdot})^2$	$SS/(Jn-1)$	

- For school A, effect is 28.4 with se 14.9
- For school A, effect is 7.9 with se 4.2

Nonhierarchical Approaches

Separate Estimates

- Consider each estimate θ_j separately.

$$\bar{y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

Pooled Estimates

- Consider a single common effect

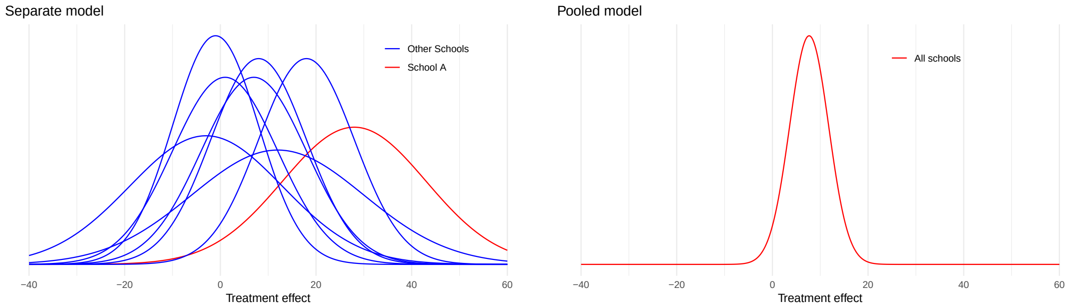
$$\bar{y}_{\cdot\cdot} = \frac{1}{\sum_{j=1}^{J} n_j}\sum_{i,j} y_{ij} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}\bar{y}_{\cdot j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}$$

	df	SS	MS	$\mathrm{E}(\mathrm{MS}\mid\sigma^2,\tau)$
Between Groups	$J-1$	$\sum_i\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot j})^2$	$SS/(J-1)$	$n\tau^2+\sigma^2$
Within Groups	$J(n-1)$	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot j})^2$	$SS/J(n-1)$	σ^2
Total	$Jn-1$	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot\cdot})^2$	$SS/(Jn-1)$	

- For school A, effect is 28.4 with se 14.9
- $P(\theta_1 > 28.4) = \frac{1}{2}$?

- For school A, effect is 7.9 with se 4.2
- $P(\theta_1 < 7.9) = \frac{1}{2}$?

The ETS Dataset



Hierarchical Approaches

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}$$

- The separate estimate is posterior mean if J values θ_j have independent uniform prior density on $(-\infty, \infty)$.

Hierarchical Approaches

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}$$

- The separate estimate is posterior mean if J values θ_j have independent uniform prior density on $(-\infty, \infty)$.
- The pooled estimate is posterior mean if J values θ_j are restricted to be equal, with uniform prior on common θ.

Hierarchical Approaches

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}$$

- The separate estimate is posterior mean if J values θ_j have independent uniform prior density on $(-\infty, \infty)$.
- The pooled estimate is posterior mean if J values θ_j are restricted to be equal, with uniform prior on common θ.
- The weighted combination is posterior mean if J values θ_j have i.i.d normal prior.

Hierarchical Approaches

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}$$

- The separate estimate is posterior mean if J values θ_j have independent uniform prior density on $(-\infty, \infty)$.
- The pooled estimate is posterior mean if J values θ_j are restricted to be equal, with uniform prior on common θ.
- The weighted combination is posterior mean if J values θ_j have i.i.d normal prior.

$$p(\theta_1, \cdots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} N(\theta_j|\mu, \tau)$$

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

Hierarchical Approaches

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}$$

- The separate estimate is posterior mean if J values θ_j have independent uniform prior density on $(-\infty, \infty)$.
- The pooled estimate is posterior mean if J values θ_j are restricted to be equal, with uniform prior on common θ.
- The weighted combination is posterior mean if J values θ_j have i.i.d normal prior.

$$p(\theta_1, \cdots, \theta_J | \mu, \tau) = \prod_{j=1}^{J} N(\theta_j | \mu, \tau)$$

$$p(\mu, \tau) = p(\mu | \tau) p(\tau) \propto p(\tau)$$

- θ_j's are conditionally independent given (μ, τ).

Hierarchical Approaches

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}$$

- The separate estimate is posterior mean if J values θ_j have independent uniform prior density on $(-\infty, \infty)$.
- The pooled estimate is posterior mean if J values θ_j are restricted to be equal, with uniform prior on common θ.
- The weighted combination is posterior mean if J values θ_j have i.i.d normal prior.

$$p(\theta_1, \cdots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} N(\theta_j|\mu, \tau)$$

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

- θ_j's are conditionally independent given (μ, τ).
- Assign non-informative uniform hyperprior to μ given τ. We can afford to be vague because J experiments are highly informative about μ.

Hierarchical Approaches: Joint and Conditional Posterior Distribution

- Joint Posterior Distribution:

$$p(\theta, \mu, \tau | y) \propto p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta)$$
$$\propto p(\mu, \tau) \prod_{j=1}^{J} N(\theta_j | \mu, \tau^2) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \theta_j, \sigma_j^2)$$

- Conditional Posterior Distribution: θ_j are conditionally independent, so the conditional posterior has J components. Each of them is a normal so

$$\bar{y}_{\cdot j} | \theta_i \sim N(\theta_i, \sigma_j^2); \theta_i | (\mu, \tau) \sim N(\mu, \tau^2) \implies \theta_j | \mu, \tau, y \sim N(\hat{\theta}_j, V_j)$$

where

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2} \bar{y}_{\cdot j} + \frac{1}{\tau^2} \mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \text{ and } V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

Hierarchical Approaches: Marginal Posterior Distribution

Fully Bayesian Treatment for $p(\mu, \tau | y)$:

1. Brute Force:
$$p(\mu, \tau | y) = \int p(\theta, \mu, \tau | y) d\theta$$
2. Analytic Solution:
$$p(\mu, \tau | y) = \frac{p(\theta, \mu, \tau | y)}{p(\theta | y)}$$
3. There is a different solution for Hierarchical Normal Model
$$p(\mu, \tau | y) \propto p(\mu, \tau) p(y | \mu, \tau)$$

- $p(y|\mu, \tau)$ could be written in closed form: $\bar{y}_{\cdot j} | \mu, \tau \sim N(\mu, \sigma_j^2 + \tau^2)$

$$\mathrm{E}(e^{it\bar{y}_{\cdot j}} | \mu, \sigma) = \mathrm{E}(\mathrm{E}(e^{it\bar{y}_{\cdot j}} | \theta_j) | \mu, \sigma) = \mathrm{E}(e^{it\theta_j} | \mu, \sigma) e^{-\frac{1}{2}\sigma_j^2} = e^{it\mu - \frac{1}{2}(\sigma_j^2 + \tau^2)}$$

Hierarchical Approaches: Posterior distribution of μ given τ. $p(\mu|\tau, y)$

- From previous page

$$p(\mu, \tau|y) \propto p(\tau)p(\mu|\tau)\prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\mu, \sigma_j^2 + \tau^2)$$

- Given τ fixed, and $p(\mu|\tau) \propto 1$, the $\log p(\mu, \tau|y)$ is quadratic in μ. This implies $p(\mu|\tau, y)$ must be normal.

$$\mu|\tau, y \sim N(\hat{\mu}, V_\mu)$$

where

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}\bar{y}_{\cdot j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}} \text{ and } V_\mu = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$

Hierarchical Approaches: Posterior distribution of τ. $p(\tau|y)$

$$
\begin{aligned}
p(\tau|y) &= \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \propto \frac{p(\tau)p(\mu|\tau)p(y|\mu,\tau)}{p(\mu|\tau,y)} \\
&\propto \frac{p(\tau)p(\mu|\tau)\prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\mu,\sigma_j^2+\tau^2)}{N(\hat{\mu},V_\mu)} \\
&\propto p(\tau)\frac{\prod_j(\sigma_j^2+\tau^2)^{-1/2}\exp\left[-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{y}_{\cdot,j}-\mu)^2\right]}{V_\mu^{-1/2}\exp\left[-\frac{1}{2V_\mu}(\mu-\hat{\mu})^2\right]} \\
&\propto p(\tau)\frac{\prod_j(\sigma_j^2+\tau^2)^{-1/2}\exp\left[-\frac{1}{2(\sigma_j^2+\tau^2)}(\mu-\hat{\mu}+\hat{\mu}-\bar{y}_{\cdot,j})^2\right]}{V_\mu^{-1/2}\exp\left[-\frac{1}{2V_\mu}(\mu-\hat{\mu})^2\right]} \\
&\propto p(\tau)V_\mu^{1/2}\prod_{i=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left(-\frac{(\bar{y}_{\cdot j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right)
\end{aligned}
$$

Hierarchical Approaches: Posterior distribution of τ. $p(\tau|y)$

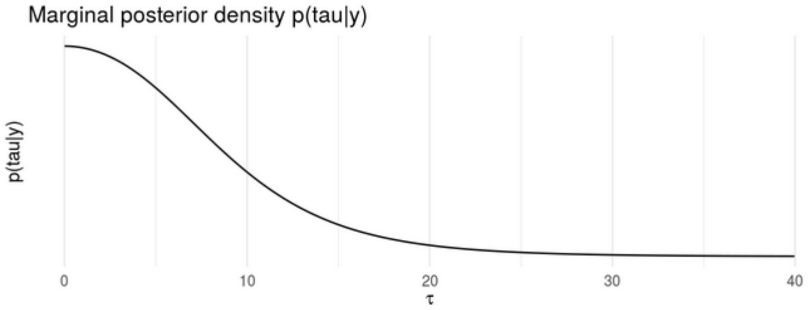
$$-\sum_j \frac{1}{(\sigma_j^2+\tau^2)}(\mu-\hat{\mu}+\hat{\mu}-\bar{y}_{\cdot,j})^2+\frac{1}{V_\mu}(\mu-\hat{\mu})^2$$

$$=\overbrace{-\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})^2}{(\sigma_j^2+\tau^2)}}^{keep}-\sum_j \frac{(\mu-\hat{\mu})^2}{(\sigma_j^2+\tau^2)}+\sum_j \frac{2(\hat{\mu}-\bar{y}_{\cdot,j})(\mu-\hat{\mu})}{(\sigma_j^2+\tau^2)}+\frac{1}{V_\mu}(\mu-\hat{\mu})^2$$

$$=\overbrace{-\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})^2}{(\sigma_j^2+\tau^2)}}^{keep}+2(\mu-\hat{\mu})\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})}{(\sigma_j^2+\tau^2)}$$

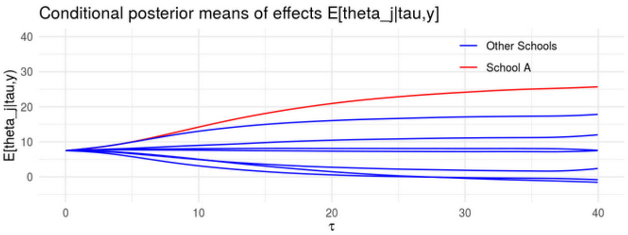
$$=\overbrace{-\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})^2}{(\sigma_j^2+\tau^2)}}^{keep}+2(\mu-\hat{\mu})(\hat{\mu}V_\mu^{-1}-\hat{\mu}V_\mu^{-1})=-\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})^2}{(\sigma_j^2+\tau^2)}$$

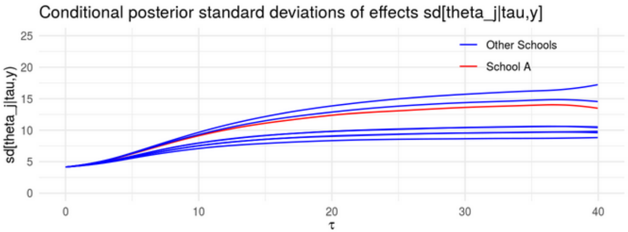
The ETS Dataset: Evaluate $p(\tau|y)$ on a Grid



- Values of τ near zero are most plausible;
- Values of τ larger than 10 are less than half as likely as $\tau = 0$, and $P(\tau > 25) \approx 0$.

The ETS Dataset: $p(\theta_j|\tau, y) = \int p(\theta_j|\mu, \tau, y)p(\mu|\tau, y)\mathrm{d}\mu$





- For most of the likely values of τ, the estimated effects are relatively close together;
- As τ increases, the population distribution allows the eight effects to be more different from each other, and hence the posterior uncertainty in each individual θ_j increases.

The ETS Dataset:Conclusion

- v.s. Pooled
  - Too much pulling together of the estimates in the eight schools; - τ is on the boundary of its parameter space.
- v.s. Seperate
  - Ordering of the effects in the eight schools is essentially the same as would be obtained by the eight separate estimates.
  - Bayesian probability that the effect in school A is as large as 28 points is less than 10%, which is substantially less than the 50% probability based on the separate estimate for school A.
- Hierarchical model is flexible enough to adapt to the data, thereby providing posterior inferences that account for the partial pooling as well as the uncertainty in the hyperparameters.