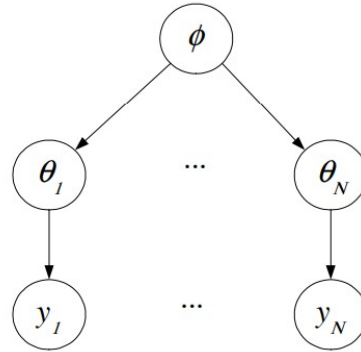


Normal Hierarchical Models



Statistics Foundations

Jorge III Altamirano Astorga
Ángel Rafael Ortega Ramirez

Motivation – Introduction

- Many stat applications involve hierarchical data, so hierarchical models are more appropriate, as it's possible to structure some dependence.
- Having insufficient parameters, they tend to overfit.
- Can be used for “meta-analysis”: used for research in order to understand a relationship between different related experiments.

Hierarchical Models

- Some authors coin the term *Empirical Bayes* to the analysis using the data to estimate prior parameters.
- Exchangeability: if no information is given to distinguish any of the θ_j . Then, no order or grouping of the parameters can be made. Ignorance of this info implies exchangeability.

$$p(\theta|\phi) = \prod_{j=1}^J p(\theta_j|\phi)$$

$$p(\theta) = \int [\prod_{j=1}^J p(\theta_j|\phi)] p(\phi) d\phi$$

Hierarchical Models

Joint Posterior Distribution:

$$p(\phi, \theta) = p(\phi)p(\theta|\phi)$$

Hyperprior Distribution for ϕ :

$$p(\phi, \theta|y) \propto p(\phi, \theta)p(y|\phi, \theta) = p(\phi, \theta)p(y|\theta)$$

- May use a diffuse distribution, if little is known.
- Should result in a posterior dist. that is proper.
- Should at least constrain the hyper params into a finite region.

Normal Hierarchical Models

Normal hierarchical models

Normal hierarchical models

Suppose we have the following model

$$\begin{aligned}y_{ij} &\stackrel{\text{ind}}{\sim} N(\theta_j, \sigma^2) \\ \theta_j &\stackrel{\text{iid}}{\sim} N(\mu, \tau^2) \\ p(\mu, \tau) &\propto I(\tau > 0)\end{aligned}$$

with $i = 1, \dots, n_j$ and $j = 1, \dots, J$. This is a normal hierarchical model.

For the moment, we assume σ^2 is known for computational reasons.

Hierarchical Normal Distributions

- The marginal posterior distribution of ϕ can be computed algebraically using the conditional probability formula.

$$p(\phi|y) = \frac{p(\phi, \theta|y)}{p(\theta|\phi, y)}$$

- The denominator has a normalizing factor that depends on ϕ and θ : this is the difficult part, as it depends on ϕ , y .
- Care must be taken to make sure that the proportionality constant (denominator) is actually a constant.
- Many times, a conjugate hierarchical scheme assumes normal sampling and normally distributed latent effects.

Estimating an Exchangeable Set of Params for a Normal Model

- We will show a simple normal hierarchical model: one way normal random effects model.
- Different mean for each group or experiment.
- Known variance, this assumption can be an adequate approximation at the sampling level.
- Data: $y_i | \theta_j \sim N(\theta_j, \sigma^2)$ for $i = 1, \dots, n_j; j = 1, \dots, J$
Likelihood : $\bar{y}_{.j} | \theta_j \sim N(\theta_j, \sigma_j^2)$
Analysis of variance : $\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$

Estimating an Exchangeable Set of Params for a Normal Model

- Pooled estimate: $\overline{y_{..}} = \frac{\sum_{j=1}^J \frac{\overline{y_{.j}}}{\sigma_j^2}}{\sum_{j=1}^J \frac{1}{\sigma_j^2}}$
- Traditionally, the analysis was to test differences among means. Choosing between the lesser.
- Alternatively we can use both: weighted combination: $\hat{\theta}_j = \lambda_j \overline{y_{.j}} + (1 - \lambda_j) \overline{y_{..}}$

Normal Hierarchical Model

- $p(\theta_1, \dots, \theta_J | \mu, \tau) = \prod_{j=1}^J N(\theta_j | \mu, \tau^2)$
 $p(\theta_1, \dots, \theta_J) = \int \prod_{j=1}^J [N(\theta_j | \mu, \tau^2)] p(\mu, \tau) d(\mu, \tau)$
- We can assign a noninformative uniform hyperprior distribution to μ , given τ :

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

Joint Posterior Distribution

- Combining the sampling distribution for the observable y_{ij} we can express:

$$\begin{aligned} p(\theta, \mu, \tau | y) &\propto p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta) \\ &\propto p(\mu, \tau) \prod_{j=1}^J N(\theta_j | \mu, \tau^2) \prod_{j=1}^J N(\overline{y}_{.j} | \theta_j, \sigma^2) \end{aligned}$$

- We can say that θ parameters are independent of the prior distribution, then:

$$\begin{aligned} \theta_j | \mu, \tau, y &\sim N(\hat{\theta}_j, V_j) \dots \text{being proper} \\ \text{where, } \hat{\theta}_j &= \frac{\overline{y}_{.j} \sigma^{-2} + \mu \tau^{-2}}{\sigma_j^{-2} + \tau^{-2}}, V_j = \frac{1}{\sigma_j^{-2} + \tau^{-2}} \end{aligned}$$

Marginal Posterior Distribution (Hyperparameters)

- Prev slide is only a partial solution. As μ and τ are unknown.
- In the normal models the marginal likelihood has a simple form (which is not the case in other distributions).

$$\begin{aligned}\overline{y_{\cdot j}}|\mu, \tau &\sim N(\mu, \sigma_j^2 + \tau^2) \\ p(\mu, \tau|y) &\propto p(\mu, \tau) \prod_{j=1}^J N(\overline{y_{\cdot j}}|\mu, \sigma_j^2 + \tau^2) \\ \hat{\mu} &= \frac{\sum_{j=1}^J \frac{\overline{y_{\cdot j}}}{\sigma_j^2 + \tau^2}}{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}}, V_{\mu}^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}\end{aligned}$$

Posterior distribution of τ

- We can now get it:
$$p(\tau|y) = \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)}$$
- This holds true as
$$\propto \frac{p(\tau) \prod_{j=1}^J N(\overline{y_{\cdot j}} | \mu \sigma_j^2 + \tau^2)}{N(\mu | \hat{\mu}, V_\mu)}$$
 any value of μ , so all the factors of μ must cancel when simplification is done.
- If we set μ to $\hat{\mu}$

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

Prior distribution of τ

- Prior distribution for τ should be assigned: a diffuse noninf. prior will be used for convenience.
- It should be a finite integral: $p(\tau) \propto 1$
- Other priors can be defined:
 - $p(\log \tau) \propto 1$
 - Inverse chi-squared: being a natural choice for variance parameters.

Posterior Predictive Distributions

- There are 2 scenarios
 - Taking the future data from current batches
$$\theta = (\theta_1, \dots, \theta_j)$$
 - Future data from future batches (y_{new})
$$\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_{\tilde{j}})$$
 - Steps to follow in this scenario:
 - (i) Draw (μ, τ) from the posterior, (ii) draw \hat{J} new params from the population distribution $p(\tilde{\theta}_J | \mu, \tau)$, (iii) draw \hat{y} given $\tilde{\theta}$ from the distribution

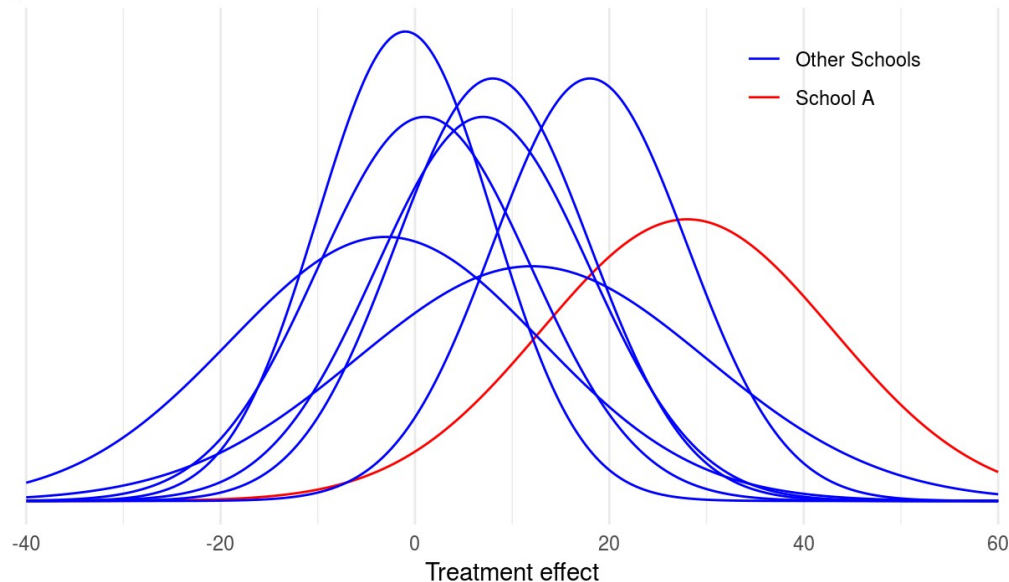
Example

- SAT-Verbal is applied to 8 different schools with a coaching program.
 - Variable of interest: score, with values 200-800, mean 500, standard deviation 100

Example

- Separate estimates: It's statistically difficult to distinguish between experiments: yielding 95% posterior intervals overlapping.

Separate model

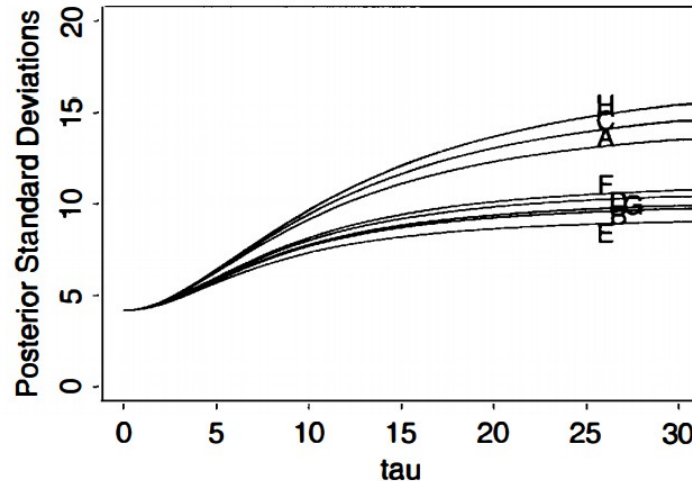


Example:

- Overlapping in the normal non pooled simple model suggests that we're estimating the same quantity.
- Pooled estimate, in contrast, hypothesize that all experiments have the same effect and produce independent estimates of this common effect.

Example

- Anyway, the posterior mode of τ is on the boundary of its parameter space.
- The same for the joint posterior modal estimate.



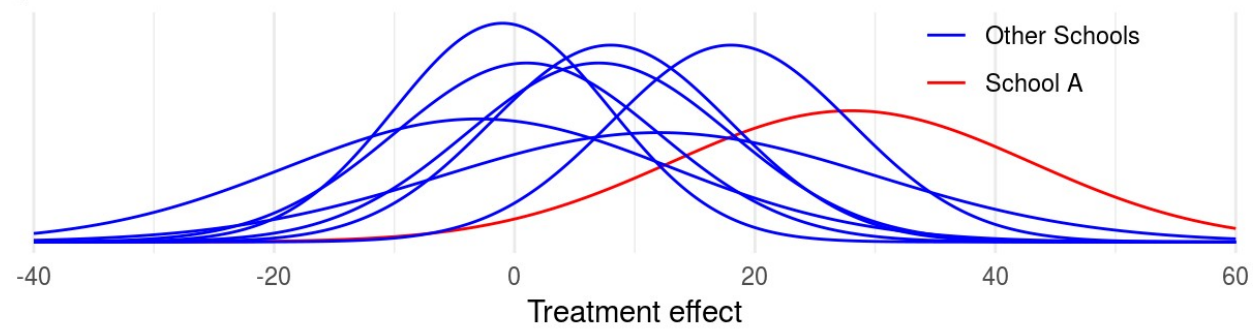
Meta-Analysis

- It's increasingly popular to summarize and integrate the findings of research studies.
- It's a method for combining several parallel data sources.
- We introduce 0 subscripts for control groups, and 1 in treatment groups.

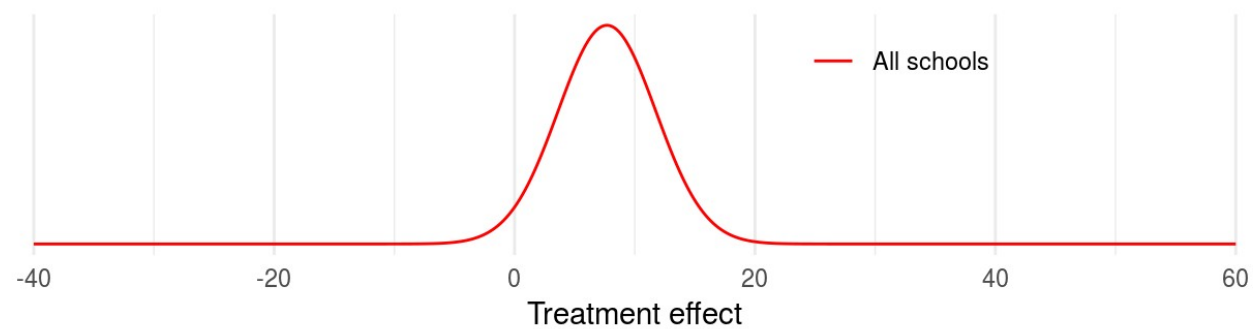
Example: Hierarchical Normal Model

- $\sigma_j^2 = \frac{1}{y_{1j}} + \frac{1}{n_{1j}-y_{1j}} + \frac{1}{y_{0j}} + \frac{1}{n_{0j}-y_{0j}}$ *Appr sampling variance*
 $y_j = \log\left(\frac{y_{1j}}{n_{1j}-y_{1j}}\right) - \log\frac{y_{0j}}{n_{0j}-y_{0j}}$ *empirical logits*
 $y_j|\theta_j, \sigma_j^2 \sim N(\theta_j, \sigma_j^2), j = 1, \dots, J$
- This method has a marginal posterior that peaks at nonzero value, which is plausible.
- The Expected value of μ is shrinked, compared with the non-hierarchical model. But the variance is accounted for, as it's uncertain in the estimation.

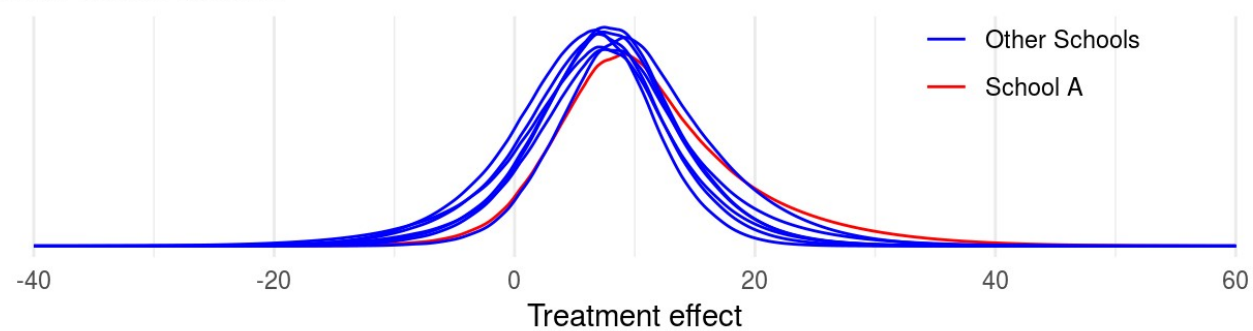
Separate model



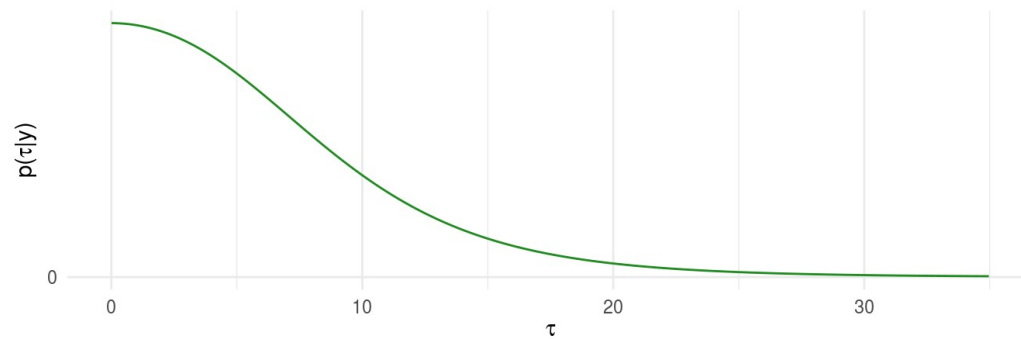
Pooled model



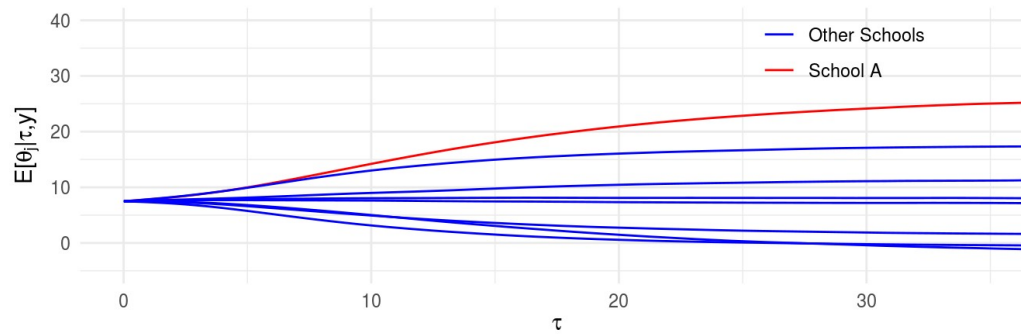
Hierarchical model



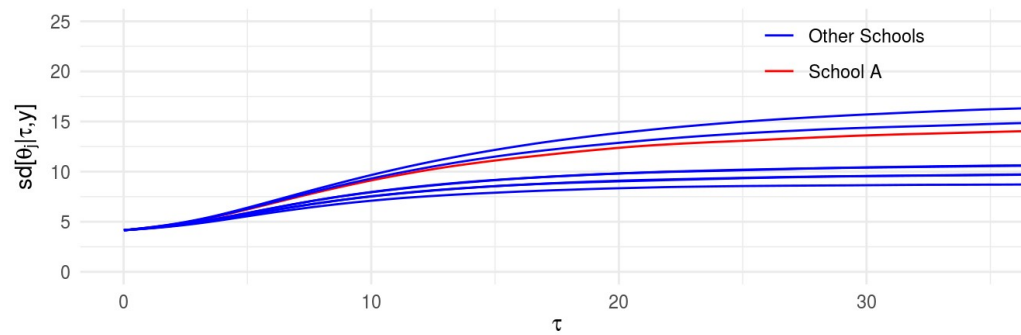
Marginal posterior $p(\tau|y)$



Conditional means $E[\theta_j|\tau,y]$



Conditional standard deviations $sd[\theta_j|\tau,y]$



BUGS Example

- <https://philwebsurfer.github.io/fundstats/>

References

- Carlin, J., Gelman, A., & Stern, H., et al. (2003). Bayesian data analysis. Chapman and Hall/CRC.
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). The Bugs Book: A Practical Introduction to Bayesian Analysis (1st ed.).
- Goldstein, H. (2011). Multilevel statistical models. Oxford: Wiley.
- Congdon, P. (2010). Applied Bayesian hierarchical methods. Boca Raton: CRC Press.