Hierarchical Models

Consider the example presented in the textbook about combining information from educational testing experiments in eight schools.

A study was performed to analyze the effects of special coaching programs on SAT-V scores in 8 high schools. The observed effects of special preparation are estimates based on separate analyses for the eight school experiments. The effects, are labeled as y_j. Over 30 students were tested on each school.

School	A	B	C	D	E	F	G	H
y_j	28.39	7.94	-2.75	6.82	-0.64	0.63	18.01	12.16
σ_j	14.9	10.2	16.3	11.0	9.4	11.4	10.4	17.6

A Statistical Model

We have $J = 8$ independent experiments, each one is performed to estimate θ_j, the unobserved true effect, from n_j independent data points y_{ij}. We assume normality and write

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

with σ^2 known. Then let

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

What posterior estimates might be reasonable for $\theta = (\theta_1, \ldots, \theta_J)$?

ANOVA Analysis

The traditional classical statistics approach is to perform an analysis of variance (ANOVA) to test for differences among the means. The resulting inference is that the means are either all equal or there are some differences between them.

Variability	DF	SS	MS
Between	$J-1$	$\sum_j (\overline{y}_{.j} - \overline{y}_{..})^2$	$SS/(J-1)$
Within	$J(n-1)$	$\sum_{ij} (y_{ij} - \overline{y}_{.j})^2$	$SS/(J(n-1))$
Total	$Jn-1$	$\sum_{ij} (y_{ij} - \overline{y}_{..})^2$	$SS/(Jn-1)$

From this table we obtain the ratio MSB/MSW. If this is significantly greater than 1, then the data favor $\hat{\theta}_j = \overline{y}_{.j}$. If not, then $\hat{\theta}_j = \overline{y}_{..}$, for all j.

Bayesian Analysis

Then, one solution indicates no pooling and the other indicates total pooling. What would a Bayesian do?

- We obtain $\hat{\theta}_j = \bar{y}_{.j}$ as the posterior mean if the J values of θ_j have independent uniform priors on $(-\infty, \infty)$.

- We obtain $\hat{\theta} = \bar{y}_{..}$ if all θ_j are assumed to be equal (to θ), and we use a uniform prior on θ.

An alternative would be to obtain a weighted combination, $\hat{\theta}_j = \lambda_j \bar{y}_{.j} + (1 - \lambda_j)\bar{y}_{..}$, with $0 < \lambda_j < 1$. What kind of priors produce these estimates?

- $\hat{\theta}_j = \lambda_j \bar{y}_{.j} + (1 - \lambda_j)\bar{y}_{..}$, is the posterior mean if the θ_j's have normal iid priors.

Pooling or not?

Inferences based on considering each school separately lead to posterior probability intervals for θ_j symmetric and centered around $\bar{y}_{.j}$. This, for example, implies that the effect in the first school has 50% chance of being above 28.39.

If we assume that all effects are equal and we obtain the pooled estimate, then the common effect is 7.9, with a standard error of 4.9. In this case the probability that the effect in the first school is below 7.9 is 50%.

These are very contrasting conclusions obtained from different assumptions regarding the model for the true underlying effects.

The Hierarchical Model

Assuming that σ_j^2 is known for each j, we consider

$$\bar{y}_{.j}|\theta_j \sim N(\theta_j, \sigma_j^2), \quad \theta_j|\mu, \tau^2 \sim N(\mu, \tau^2), \quad p(\mu|\tau) \propto 1, \;\; p(\tau) \propto 1$$

Caution: if you are going to use improper priors you have to make sure that your posterior is proper.

- *Prior for $\mu|\tau$:* the uniform prior is usually reasonable because the combined data of the J experiments are usually very informative on μ.
- *Prior for τ:* using $p(\tau) \propto 1$ yields a proper posterior in this case (if $J > 2$). However, the usual 'non-informative' prior $p(\log \tau) \propto 1$ yields an improper posterior!

The joint posterior is

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

This can be explored using simulations drawn from

$$p(\theta,\mu,\tau|y) \propto p(\tau|y)p(\mu|\tau,y)p(\theta|\mu,\tau,y)$$

The last factor is *the conditional posterior for θ*

$$\theta_j|\mu,\tau,y \sim N(\hat{\theta}_j, V_j)$$

where,

$$\hat{\theta_j} = \frac{\bar{y}_{j.}/\sigma_j^2 + \mu/\tau^2}{1/\sigma_j^2 + 1/\tau^2}, \quad V_j = \frac{1}{1/\sigma_j^2 + 1/\tau^2}$$

Integrating θ_j out, we obtain *the marginal posteriors for μ and τ*

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

and we observe that

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

where

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

with

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} \bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}}, \quad V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}$$

The posterior $p(\tau|y)$ can be obtained as follows

$$\begin{aligned} p(\tau|y) &= \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \\ &\propto \frac{p(\tau)\prod_{j=1}^{J} N(\bar{y}_{.j}|\mu,\sigma_j^2+\tau^2)}{N(\mu|\hat{\mu},V_\mu)} \end{aligned}$$

Since this holds for any value of μ, we can write

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

$$\propto p(\tau)V_\mu^{1/2}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left(-\frac{(\bar{y}_{.j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right)$$

The Posterior Predictive

- To obtain future data $\tilde{y}$ from the current set of batches with mean $\theta = (\theta_1, \ldots, \theta_J)$, we first obtain a draw from $p(\theta, \mu, \tau | y)$ and then draw $\tilde{y}$ using $\tilde{y}_{ij} | \theta_j \sim N(\theta_j, \sigma_j^2)$
- To obtain posterior predictive simulations of new data $\tilde{y}$ from $\tilde{J}$ new groups
  - draw (μ, τ) from $p(\mu, \tau | y)$
  - draw $\tilde{J}$ parameters from $p(\tilde{\theta}_j | \mu, \tau)$
  - specify $\tilde{n}_j$ and draw $\tilde{y}$ from $\tilde{y}_{ij} | \tilde{\theta}_j \sim N(\tilde{\theta}_j, \sigma^2)$

R Code

```
# output: one sample from p(theta | mu, tau, y)
conditional.theta_function(ybar,mu,tau,sigma){
  theta=rep(0,nschools)
  theta.hat=rep(0,nschools)
  V.hat=rep(0,nschools)
  for(j in 1:nschools){
    V.hat[j]=1/(1/sigma[j]^2+1/(tau^2))
    theta.hat[j]=(ybar[j]/sigma[j]^2+mu/tau^2)*V.hat[j]
    theta[j]=rnorm(1,theta.hat[j],sqrt(V.hat[j]))
  }
  theta
}
```

R Code

```
# output: nsample samples from p(mu | tau, y)
sample.mar.mu=function(ybar, tau, sigma,nsample){
  V.mu.inv=sum(1/(sigma^2+tau^2))
  mu.hat=sum((1/(sigma^2+tau^2))*ybar)/V.mu.inv
  mu.sample=rnorm(nsample,mu.hat,sqrt(1/V.mu.inv))
  mu.sample
}
# evaluates p(tau | y)
marginal.tau=function(ybar,tau,sigma){
  V.mu.inv=sum(1/(sigma^2+tau^2))
  mu.hat=sum((1/(sigma^2+tau^2))*ybar)/V.mu.inv
  eval=exp(-(ybar-mu.hat)^2/(2*(sigma^2+tau^2)))
  eval=eval/sqrt(sigma^2+tau^2)
  eval=sqrt(1/V.mu.inv)*prod(eval)
  eval }
```

R Code

```
########### Main program ###################
# Read data. Data file sat.score of the form
# School Treat.effect sd.effect
# A 28.39 14.9 ...
sat.scores=read.table(’sat.scores’,header=TRUE)
ybar=sat.scores$Treat.effect
nschools=length(ybar)
sigma=sat.scores$sd.effect
# Grid to evaluate p(tau |y)
x.tau=seq(0.00001,40,length=1000)
# evaluate p(tau |y) at 1000 points in the
#interval [0.00001,40]
post.tau=apply(t(x.tau),2,marginal.tau,
ybar=ybar, sigma=sigma)
```

R Code

```
# draw 200 samples from p(tau |y)
sample.tau=sample(x.tau,200,replace=TRUE,
prob=post.tau)
# draw 200 samples from p(mu | tau, y)
sample.mu=apply(t(sample.tau),2,sample.mar.mu,
ybar=ybar, sigma=sigma,nsample=1)
# draw 200 samples from p(theta | mu, tau, y)
sample.theta=matrix(0,ncol=nschools,nrow=200)
for (i in 1:200){
sample.theta[i,]=conditional.theta(ybar,
sample.mu[i], sample.tau[i],sigma)
}
```

R Code

```
# Expected posterior means E(theta=j | tau, y)
# averaging over mu
expected.theta=matrix(0,ncol=nschools,nrow=30)
x.tau.2=seq(0.00001,30,length=30)
for (i in 1:30){
sample.mu=sample.mar.mu(ybar,x.tau.2[i],sigma,
nsample=5000)
sample.theta.2=matrix(0,ncol=nschools,nrow=5000)
for (j in 1:5000){
sample.theta.2[j,]=conditional.theta(ybar,
sample.mu[j], x.tau.2[i],sigma)
}
expected.theta[i,]=apply(sample.theta.2,2,mean)
}
```

R Code

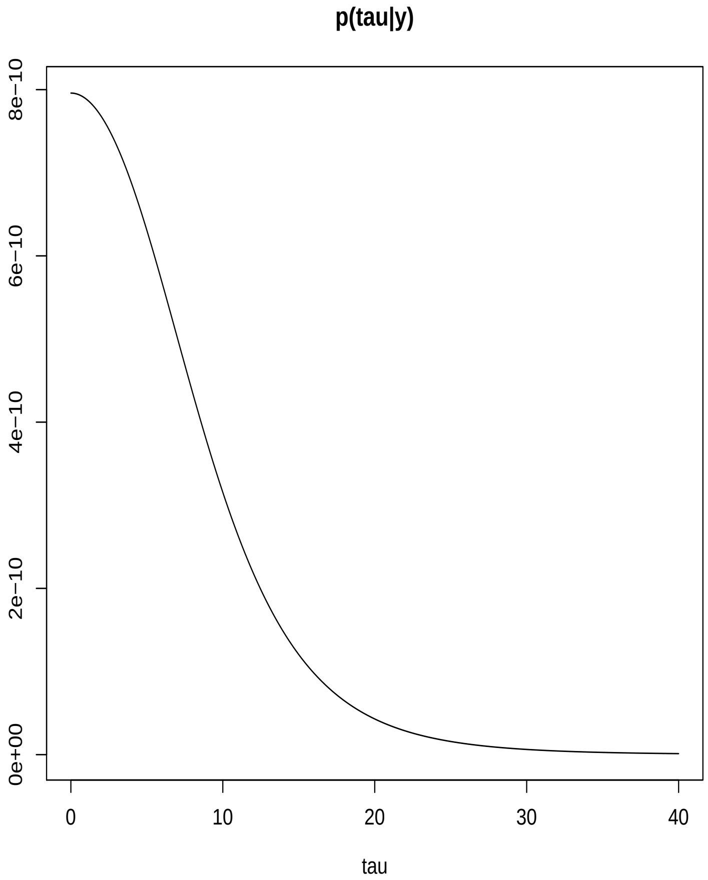
```
source(’hierarchical.r’)
plot(x.tau,post.tau,type=’l’,ylab="",xlab="tau")
title(’p(tau|y)’)
hist(sample.tau,ylab="",xlab="tau",
main="p(tau|y)",col=’lightblue’)
# 95\% P.I. for tau
sort(sample.tau)[5]
[1] 0.4804904
sort(sample.tau)[195]
[1] 19.69970
```

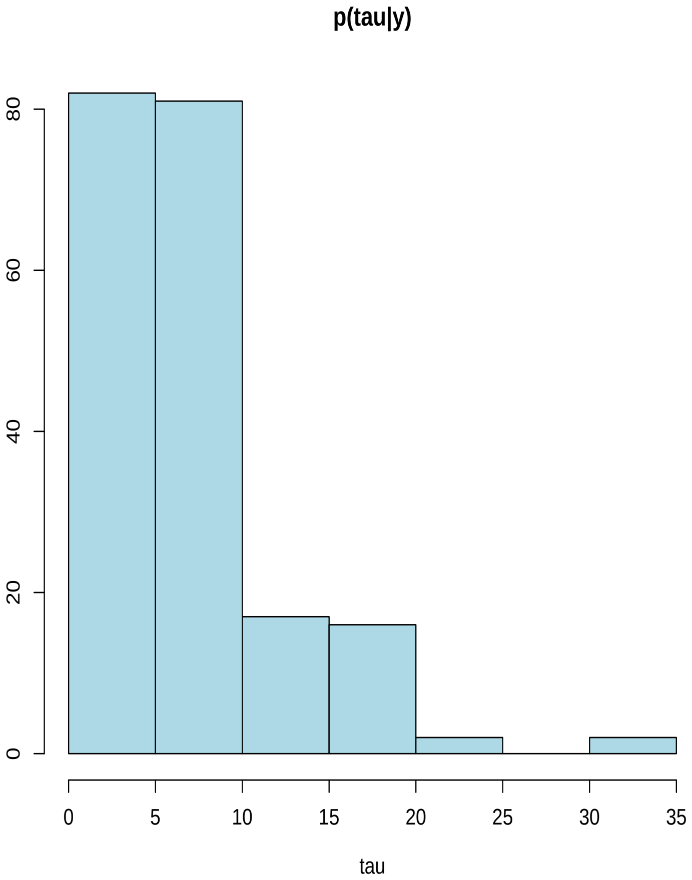
Results

Results based on 200 simulations from the posterior of all parameters produce the following estimates of the quantiles.

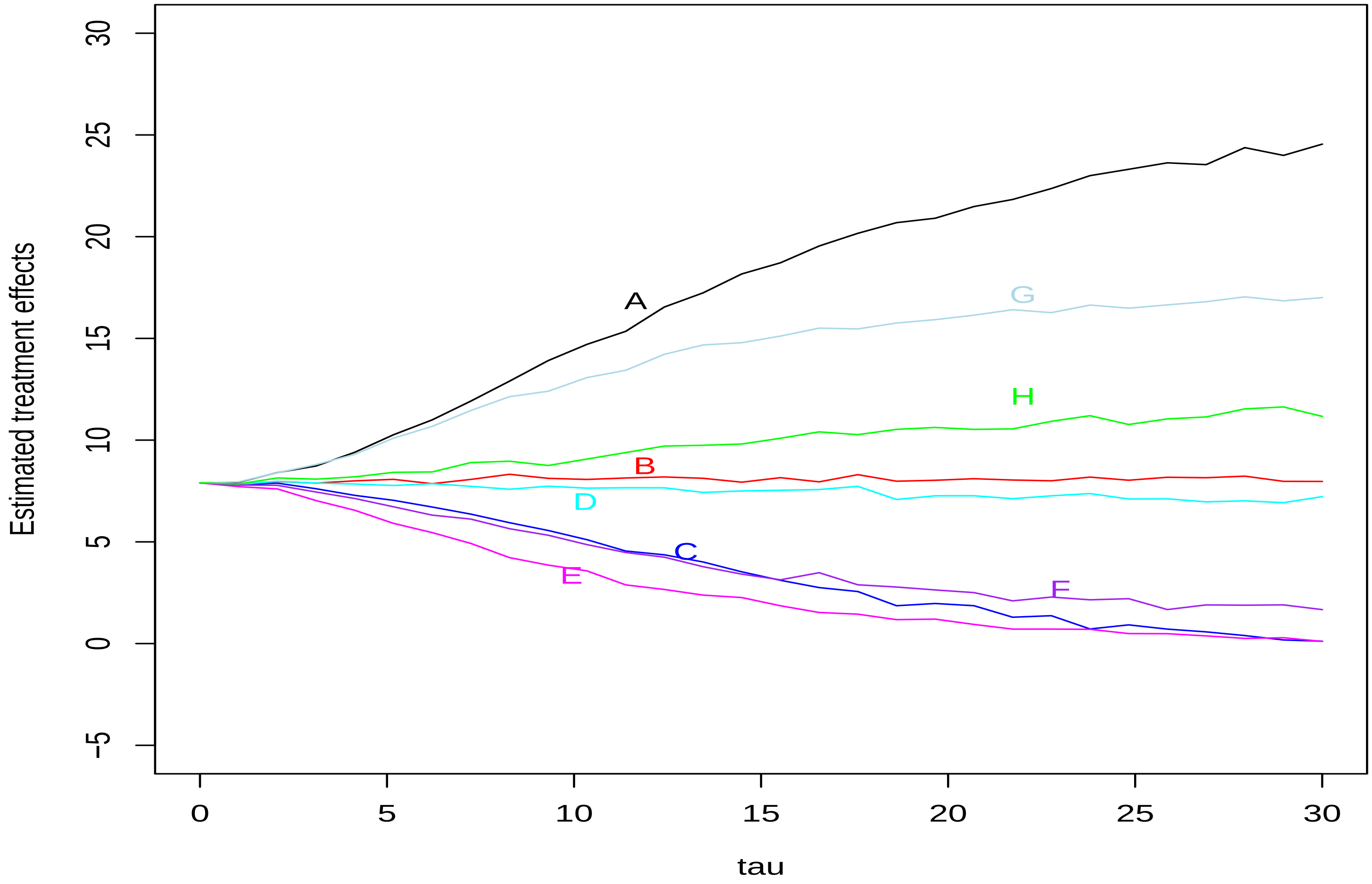
School	2.5%	50%	97.5%
A	-2.310	11.727	34.090
B	-4.387	8.077	21.811
C	-18.662	5.595	17.313
D	-6.436	7.190	23.483
E	-13.310	5.191	16.958
F	-11.441	6.176	19.357
G	-1.525	10.788	25.687
H	-6.511	8.846	24.645

Results


p(tau|y)
0e+00
2e−10
4e−10
6e−10
8e−10
0
10
20
30
40
tau


p(tau|y)
0
20
40
60
80
0
5
10
15
20
25
30
35
tau

Results


Estimated treatment effects
-5
0
5
10
15
20
25
30
0
5
10
15
20
25
30
tau
A
G
H
B
D
C
E
F