**Bayesian Hierarchical Modeling on SAT Coaching Data**

**Sample Problem 2**

(a)

The prior for is a normal distribution with 0 mean and large variance. It is close to “flat”, so it does not show strong prior belief about the value of . But the prior for is . From this prior it is likely to have small values. This gives relatively strong information. Based on the priors for and , we are relatively uncertain about the mean of , but believe that the variance of should not be large.

(b)

(c)

So the conditional posterior of given everything else is

So the conditional posterior of given everything else is

So the conditional posterior of given everything else is

(d)

R code as follows (Python code attached separately):

set.seed(1234) **# Set random seed 1234**

sigma<-c(14.9,10.2,16.3,11.0,9.4,11.4,10.4,17.6)

y<-c(28.39,7.94,-2.75,6.82,-0.64,0.63,18.01,12.16)

mu0<-rnorm(1,0,100) **# Initial value of mu**

tau\_squared0<-1/rgamma(1,0.01,0.01) **# Initial value of tau\_squared**

theta0<-numeric(8)

for (i in 1:8)

{

theta0[i]<-rnorm(1,(sigma[i]^2\*mu0+y[i]\*tau\_squared0)/(sigma[i]^2+tau\_squared0),sqrt((sigma[i]^2\*tau\_squared0)/(sigma[i]^2+tau\_squared0)))

} **# Initial values of ’s**

post\_mu<-numeric(2000)

post\_tau\_squared<-numeric(2000)

post\_theta<-matrix(data=NA,nrow=2000,ncol=8)

**# Define 2 vectors and 1 matrix to store posterior samples of mu,** **tau\_squared and ’s**

post\_mu[1]<-rnorm(1,10000\*sum(theta0)/(80000+tau\_squared0),sqrt(10000\*tau\_squared0/(80000+tau\_squared0)))

post\_tau\_squared[1]<-1/rgamma(1,4.01,0.01+sum((theta0-post\_mu[1])^2)/2)

for (j in 1:8)

{

post\_theta[1,j]<-rnorm(1,(sigma[j]^2\*post\_mu[1]+y[j]\*post\_tau\_squared[1])/(sigma[j]^2+post\_tau\_squared[1]),sqrt((sigma[j]^2\*post\_tau\_squared[1])/(sigma[j]^2+post\_tau\_squared[1])))

} **# First iteration of Gibbs sampler**

for (i in 2:2000)

{

post\_mu[i]<-rnorm(1,10000\*sum(post\_theta[i-1,])/(80000+post\_tau\_squared[i-1]),sqrt(10000\*post\_tau\_squared[i-1]/(80000+post\_tau\_squared[i-1])))

post\_tau\_squared[i]<-1/rgamma(1,4.01,0.01+sum((post\_theta[i-1,]-post\_mu[i])^2)/2)

for (j in 1:8)

{

post\_theta[i,j]<-rnorm(1,(sigma[j]^2\*post\_mu[i]+y[j]\*post\_tau\_squared[i])/(sigma[j]^2+post\_tau\_squared[i]),sqrt((sigma[j]^2\*post\_tau\_squared[i])/(sigma[j]^2+post\_tau\_squared[i])))

}

} **# Finish the whole** **Gibbs sampler, 2000 draws in total**

a<-numeric(8)

**# Define a vector to store the estimated probabilities of each being the best**

post\_est\_best<-matrix(data=0,nrow=1000,ncol=8)

**# Define a matrix to store indicators in each draw which**  **being the best**

for (k in 1001:2000) **# Use the second half of draws from** **Gibbs sampler**

{

for (j in 1:8)

{

if (post\_theta[k,j]==max(post\_theta[k,])) post\_est\_best[k-1000,j]=1

}

}

for (j in 1:8)

{

a[j]<-(sum(post\_est\_best[,j]))/1000

}

a

post\_est\_compare<-matrix(data=0,nrow=8,ncol=8)

**# Define a matrix to store estimated probabilities of pairwise comparison between schools**

for (k in 1001:2000)

{

for (m in 1:8)

{

for (n in m:8)

{

if (post\_theta[k,m]>post\_theta[k,n]) post\_est\_compare[m,n]=post\_est\_compare[m,n]+1

}

}

}

post\_est\_compare<-post\_est\_compare/1000

post\_est\_compare

**Results:**

**In the simulation, we use MCMC to generate posterior samples of schools A through H being the best respectively, and all the possible pairwise comparisons of one school being better than the other.**

**The estimated posterior probabilities of each school being the best are 0.154, 0.121, 0.115, 0.134, 0.107, 0.098, 0.140, and 0.131, from school A to H.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
| **P(best)** | **0.154** | **0.121** | **0.115** | **0.134** | **0.107** | **0.098** | **0.140** | **0.131** |

**For pairwise comparisons, results are summarized in the table below. Viewing this table as a matrix, element () (element at the cross of row and column ) of this matrix is the estimated probability of school better than school . ()**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
| **A** | **0** | **0.534** | **0.566** | **0.534** | **0.564** | **0.557** | **0.523** | **0.534** |
| **B** |  | **0** | **0.532** | **0.495** | **0.542** | **0.526** | **0.469** | **0.507** |
| **C** |  |  | **0** | **0.476** | **0.526** | **0.504** | **0.440** | **0.468** |
| **D** |  |  |  | **0** | **0.529** | **0.519** | **0.448** | **0.498** |
| **E** |  |  |  |  | **0** | **0.472** | **0.424** | **0.458** |
| **F** |  |  |  |  |  | **0** | **0.439** | **0.478** |
| **G** |  |  |  |  |  |  | **0** | **0.529** |
| **H** |  |  |  |  |  |  |  | **0** |

**All the diagonal elements of this matrix must be 0 (a school comparing to itself). Each pair of elements symmetric about the diagonal must sum up to 1. That is why we just need to present half of the matrix.**

(e)

Choose and as another two settings for .

**:**

**Estimated probability of school C being the best is 0.125.**

R codeas follows (Python code attached separately)**:**

set.seed(1234)

sigma<-c(14.9,10.2,16.3,11.0,9.4,11.4,10.4,17.6)

y<-c(28.39,7.94,-2.75,6.82,-0.64,0.63,18.01,12.16)

mu0<-rnorm(1,0,100) #normal prior

tau\_squared0<-1/rgamma(1,1,1) #Inverse Gamma prior

theta0<-numeric(8)

for (i in 1:8)

{

theta0[i]<-rnorm(1,(sigma[i]^2\*mu0+y[i]\*tau\_squared0)/(sigma[i]^2+tau\_squared0),sqrt((sigma[i]^2\*tau\_squared0)/(sigma[i]^2+tau\_squared0)))

}

post\_mu<-numeric(2000)

post\_tau\_squared<-numeric(2000)

post\_theta<-matrix(data=NA,nrow=2000,ncol=8)

post\_mu[1]<-rnorm(1,10000\*sum(theta0)/(80000+tau\_squared0),sqrt(10000\*tau\_squared0/(80000+tau\_squared0)))

post\_tau\_squared[1]<-1/rgamma(1,5,1+sum((theta0-post\_mu[1])^2)/2)

for (j in 1:8)

{

post\_theta[1,j]<-rnorm(1,(sigma[j]^2\*post\_mu[1]+y[j]\*post\_tau\_squared[1])/(sigma[j]^2+post\_tau\_squared[1]),sqrt((sigma[j]^2\*post\_tau\_squared[1])/(sigma[j]^2+post\_tau\_squared[1])))

}

for (i in 2:2000)

{

post\_mu[i]<-rnorm(1,10000\*sum(post\_theta[i-1,])/(80000+post\_tau\_squared[i-1]),sqrt(10000\*post\_tau\_squared[i-1]/(80000+post\_tau\_squared[i-1])))

post\_tau\_squared[i]<-1/rgamma(1,5,1+sum((post\_theta[i-1,]-post\_mu[i])^2)/2)

for (j in 1:8)

{

post\_theta[i,j]<-rnorm(1,(sigma[j]^2\*post\_mu[i]+y[j]\*post\_tau\_squared[i])/(sigma[j]^2+post\_tau\_squared[i]),sqrt((sigma[j]^2\*post\_tau\_squared[i])/(sigma[j]^2+post\_tau\_squared[i])))

}

}

a<-numeric(8)

post\_est\_best<-matrix(data=0,nrow=1000,ncol=8)

for (k in 1001:2000)

{

for (j in 1:8)

{

if (post\_theta[k,j]==max(post\_theta[k,])) post\_est\_best[k-1000,j]=1

}

}

for (j in 1:8)

{

a[j]<-(sum(post\_est\_best[,j]))/1000

}

a[3]

**# Everything for the code is the same as previous except change for conditional posterior of**

**:**

**Estimated probability of school C being the best is 0.126.**

R codeas follows (Python code attached separately)**:**

set.seed(1234)

sigma<-c(14.9,10.2,16.3,11.0,9.4,11.4,10.4,17.6)

y<-c(28.39,7.94,-2.75,6.82,-0.64,0.63,18.01,12.16)

mu0<-rnorm(1,0,100)

tau\_squared0<-1/rgamma(1,10,10)

theta0<-numeric(8)

for (i in 1:8)

{

theta0[i]<-rnorm(1,(sigma[i]^2\*mu0+y[i]\*tau\_squared0)/(sigma[i]^2+tau\_squared0),sqrt((sigma[i]^2\*tau\_squared0)/(sigma[i]^2+tau\_squared0)))

}

post\_mu<-numeric(2000)

post\_tau\_squared<-numeric(2000)

post\_theta<-matrix(data=NA,nrow=2000,ncol=8)

post\_mu[1]<-rnorm(1,10000\*sum(theta0)/(80000+tau\_squared0),sqrt(10000\*tau\_squared0/(80000+tau\_squared0)))

post\_tau\_squared[1]<-1/rgamma(1,14,10+sum((theta0-post\_mu[1])^2)/2)

for (j in 1:8)

{

post\_theta[1,j]<-rnorm(1,(sigma[j]^2\*post\_mu[1]+y[j]\*post\_tau\_squared[1])/(sigma[j]^2+post\_tau\_squared[1]),sqrt((sigma[j]^2\*post\_tau\_squared[1])/(sigma[j]^2+post\_tau\_squared[1])))

}

for (i in 2:2000)

{

post\_mu[i]<-rnorm(1,10000\*sum(post\_theta[i-1,])/(80000+post\_tau\_squared[i-1]),sqrt(10000\*post\_tau\_squared[i-1]/(80000+post\_tau\_squared[i-1])))

post\_tau\_squared[i]<-1/rgamma(1,14,10+sum((post\_theta[i-1,]-post\_mu[i])^2)/2)

for (j in 1:8)

{

post\_theta[i,j]<-rnorm(1,(sigma[j]^2\*post\_mu[i]+y[j]\*post\_tau\_squared[i])/(sigma[j]^2+post\_tau\_squared[i]),sqrt((sigma[j]^2\*post\_tau\_squared[i])/(sigma[j]^2+post\_tau\_squared[i])))

}

}

a<-numeric(8)

post\_est\_best<-matrix(data=0,nrow=1000,ncol=8)

for (k in 1001:2000)

{

for (j in 1:8)

{

if (post\_theta[k,j]==max(post\_theta[k,])) post\_est\_best[k-1000,j]=1

}

}

for (j in 1:8)

{

a[j]<-(sum(post\_est\_best[,j]))/1000

}

a[3]

**# Everything for the code is the same as previous except change for conditional posterior of**

**Based on the results, the posterior probability of school C being the best is not very sensitive to the choice of . For very small values, the** **estimated posterior probability tends to be small. But as values become large, there will not be great change. Actually I also tried , the estimated probability is 0.119. It becomes back to small.**

(f)

**This means that the unpooled estimate is the posterior mean if the values have independent uniform prior densities on .**

**With simple mathematical derivation, we have**

and

R codeas follows (Python code attached separately)**:**

set.seed(1234)

post.theta<-matrix(data=0,nrow=1000,ncol=8)

for (j in 1:8)

{

post.theta[,j]=rnorm(1000,y[j],sigma[j])

}

a<-numeric(8)

post.est.best<-matrix(data=0,nrow=1000,ncol=8)

for (k in 1:1000)

{

for (j in 1:8)

{

if (post.theta[k,j]==max(post.theta[k,])) post.est.best[k,j]=1

}

}

for (j in 1:8)

{

a[j]<-(sum(post.est.best[,j]))/1000

}

a

post.est.compare<-matrix(data=0,nrow=8,ncol=8)

for (k in 1:1000)

{

for (m in 1:8)

{

for (n in m:8)

{

if (post.theta[k,m]>post.theta[k,n]) post.est.compare[m,n]=post.est.compare[m,n]+1

}

}

}

post.est.compare<-post.est.compare/1000

post.est.compare

**The estimated posterior probabilities of each school being the best are 0.539, 0.027, 0.028, 0.035, 0.007, 0.010, 0.173, and 0.181 for school A to H respectively.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
| **P(best)** | **0.539** | **0.027** | **0.028** | **0.035** | **0.007** | **0.010** | **0.173** | **0.181** |

**For pairwise comparisons, results are summarized in the table below. Viewing this table as a matrix, element () (element at the cross of row and column ) of this matrix is the estimated probability of school better than school . ()**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
| **A** | **0** | **0.88** | **0.923** | **0.877** | **0.947** | **0.923** | **0.716** | **0.741** |
| **B** |  | **0** | **0.685** | **0.522** | **0.750** | **0.688** | **0.253** | **0.408** |
| **C** |  |  | **0** | **0.316** | **0.477** | **0.431** | **0.146** | **0.266** |
| **D** |  |  |  | **0** | **0.703** | **0.641** | **0.223** | **0.384** |
| **E** |  |  |  |  | **0** | **0.447** | **0.091** | **0.244** |
| **F** |  |  |  |  |  | **0** | **0.140** | **0.283** |
| **G** |  |  |  |  |  |  | **0** | **0.605** |
| **H** |  |  |  |  |  |  |  | **0** |

**All the diagonal elements of this matrix must be 0(a school comparing to itself). Each pair of elements symmetric about the diagonal must sum up to 1. The structure of the above tables is the same as that in (d).**

(g)

In a classical random-effects ANOVA, we assume the 8 treatment effects come from a distribution . We want to do the hypothesis test:

Under , pooled estimate

Chi-square statistic , p-value=0.707.

Fail to reject the null, should take the pooled estimate for ’s.

(h)

**The results from (f) are much more “absolute” while the results from (d) are relatively “mild”. I mean, for example, school A is “overwhelmingly” better than school C, E, F in (f) (Probability > 0.9), while in (d), all the schools tend to be quite close. No one can be better than any of others with an estimated probability larger than 0.6 or smaller than 0.4. In (f), school A is the best with estimated probability over a half. But in (d), the estimated probabilities of each school being the best are close to each other. We are kind of “neutral” and “cautious” and avoid having too strong conclusions.**

**Conclusions of Bayesian methods are impacted by the prior to a large degree, while conclusions of the classical method tends to be “plain” (totally driven by data). Classical method is kind of “cut with one knife” and is sometimes difficult to get further, deeper, and more sophisticated results.**

**Classical method depends totally on data (and of course, model assumptions) while Bayesian method takes more factors into account including prior belief. It is kind of using prior distributions than just assumptions to reflect information before data collection.**

**From all the above, I would prefer Bayesian method in this problem setting. Actually Bayesian inference is gaining more popularity over recent years.**