

Due: Fri, Dec, 12 in class

1. This question is a continuation of your first homework for this class.

Refer to section 1.4 (pp. 9-11) in Gelman *et al.* The question of interest is whether or not the woman is a carrier of the hemophilia gene.

Suppose the woman has 1 son, whose hemophilia status is  $y = 0$ . (This is different from the outcomes discussed in the book.)

Compute:

- (a) prior odds in favor of  $\theta = 0$  vs.  $\theta = 1$
- (b) Bayes factor in favor of  $\theta = 0$  vs.  $\theta = 1$
- (c) posterior odds in favor of  $\theta = 0$  vs.  $\theta = 1$

a. prior odds in favor the  $\theta = 0$  vs  $\theta = 1$

$$\frac{0.5}{0.5} = 1$$

To get b. and c., we need the posterior probability that  $\theta = 0$ .

$$\begin{aligned} \Pr(\theta=0|y=0) &= \frac{\Pr(\theta = 0) \Pr(y=0 | \theta = 0)}{\Pr(\theta=0) \Pr(y=0 | \theta=0) + \Pr(\theta=1) \Pr(y=0 | \theta=1)} \\ &= \frac{0.5 (1)}{0.5 (1) + 0.5(0.5)} = \frac{2}{3} \end{aligned}$$

c. So posterior odds in favor of  $\theta = 0$  vs.  $\theta = 1$  are

$$\frac{2/3}{1/3} = 2 \quad \text{and}$$

b. Bayes factor in favor of  $\theta = 0$  is  $2/1 = 2$

2. Attached is WinBUGS code and output for the Dyes example, which you know from a previous homework. I have added 4 lines to the program. Refer to the code and output to answer the following questions (just a sentence or two for each).

(a) Explain the meaning of the following new nodes:

- i. `yprcd`  
replicated datasets drawn from the posterior predictive distribution
- ii. `resid`  
absolute values of standardized residuals from real data (Note that standardization was not necessary here; if I had wanted to compare residuals to some fixed criterion based on  $N(0,1)$  distribution (e.g.  $z < 1.96$  or  $z < 2.58$ ) then the standardization would have been needed.)
- iii. `presid`  
absolute values of standardized residuals from replicated datasets
- iv. `pppv`  
For each batch  $i$ , `pppv[i] = 1` if largest absolute value of residual from real data is larger than largest absolute value of residual from replicated data, 0 otherwise. Thus the mean of `pppv[i]` is the proportion of the time that a residual in real data in batch  $i$  is more extreme than residuals in data drawn from our model.

(b) What could you learn by monitoring "pppv"?

If the mean `pppv` was close to 1 for a particular batch, it could mean an outlier in that batch. If mean `pppv` was close to 0, that would mean that the values in that batch are not as spread out as expected under our model. Either case could indicate a need for expanding the model by allowing for different variances in different batches (that is, `tau.with[i]`).

(c) Does the output suggest any problems with model fit?

No. None of the mean `pppv`'s are extremely close to either zero or 1. The GCSR textbook suggests that `pppv`'s less than 0.01 or greater than 0.99 indicate major failure of the model.

```

model
{
  for( i in 1 : batches ) {
    m[i] ~ dnorm(theta, tau.btw)
    for( j in 1 : samples ) {
      y[i , j] ~ dnorm(m[i], tau.with)
      ypred[i,j] ~ dnorm(m[i], tau.with)
      resid[i,j] <- abs(y[i,j] - m[i])
      presid[i,j] <- abs(ypred[i,j] - m[i])
    }
    large[i] <- ranked(resid[i,], samples)
    largepred[i] <- ranked(presid[i,], samples)
    pppv[i] <- step(large[i] - largepred[i])
  }
  sigma2.with <- 1 / tau.with
  sigma2.btw <- 1 / tau.btw
  tau.with ~ dgamma(0.001, 0.001)

```

```
tau.btw ~ dgamma(0.001, 0.001)
theta ~ dnorm(0.0, 1.0E-10)
}
```

```
node mean sd MC error 2.5% median 97.5% start sample
m[1] 1515.0 20.31 0.4391 1472.0 1517.0 1550.0 10001 40000
m[2] 1528.0 18.62 0.2291 1490.0 1527.0 1566.0 10001 40000
m[3] 1548.0 22.12 0.5962 1512.0 1547.0 1594.0 10001 40000
m[4] 1511.0 21.42 0.5458 1466.0 1513.0 1547.0 10001 40000
m[5] 1569.0 30.38 1.143 1517.0 1572.0 1624.0 10001 40000
m[6] 1495.0 27.22 0.953 1443.0 1494.0 1544.0 10001 40000
pppv[1] 0.561 0.4963 0.005088 0.0 1.0 1.0 10001 40000
pppv[2] 0.1466 0.3537 0.003352 0.0 0.0 1.0 10001 40000
pppv[3] 0.3195 0.4663 0.003411 0.0 0.0 1.0 10001 40000
pppv[4] 0.6383 0.4805 0.006751 0.0 1.0 1.0 10001 40000
pppv[5] 0.5282 0.4992 0.003993 0.0 1.0 1.0 10001 40000
pppv[6] 0.2517 0.434 0.004228 0.0 0.0 1.0 10001 40000
sigma2.btw 2125.0 3716.0 65.35 0.004705 1237.0 9942.0 10001 40000
sigma2.with 3083.0 1125.0 31.27 1571.0 2855.0 5842.0 10001 40000
```

```
Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes
Dbar Dhat pD DIC
Y 966.587 912.547 54.039 1020.630
total 966.587 912.547 54.039 1020.630
```

Birats

```
Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes
Dbar Dhat pD DIC
Y 969.345 919.472 49.874 1019.220
total 969.345 919.472 49.874 1019.220
```

3. Use the DIC to compare the fits of two different models for the data involving growth of baby rats. First use the model, data, and initial values exactly as given in the “Rats” example in Volume 1 of examples. Then use the model, data, and initial values as given in the “Birats” example in Volume 2. In both cases, run at least 1000 burn-in iterations before you Set the DIC. Then use output for the DIC based on at least 10000 additional iterations. Turn in the tables of DIC results for both models, and answer the following questions:

(a) What is the estimated number of free parameters in the “Rats” model? In the “Birats” model? What could explain the difference between the two estimates?

Rats: 54  
Birats: 50

The Birats model allows for correlation between rat-specific intercepts and rat-specific slopes, whereas the Rats model does not. Since more highly-correlated parameters are counted as fewer free parameters, the Birats model is shown to have fewer free parameters.

(b) Is one model strongly preferred over the other after the penalty for model complexity is taken into account? Justify your answer.

No. The DIC is very similar for both models (1020.63 vs. 1019.22). It is somewhat surprising to me that the actual fit of the Rats model looks better than that of Birats (smaller Dbar and Dhat for Rats).

Rats; univariate