MAS8303 Modern Bayesian Inference
Part 2

TEST: Solution

Semester 1, 2012-3

Time allowed: ONE HOUR.

There are TWO questions. Answer both questions.

This is an "OPEN BOOK" test. You may use books, lecture notes etc. but you may not consult anyone other than the invigilator during the test. You may use the computer and a calculator. Statistical tables are not provided but you are allowed to use your own or to use the computer instead.

Write your answers in the spaces provided on the question paper. If you wish to send me any graphs or other files then please email them to me at the following address.

`malcolm.farrow@newcastle.ac.uk`

There are 25 marks available in total.

1. The data given below are taken from Best and Walker (1964). They refer to male populations divided by age-group and smoking habits. In each group the population size and the number dying in a particular time interval are given.

Age	Nonsmokers		Cigar and pipe only		Cigarette and other		Cigarette only	
	Pop.	Deaths	Pop.	Deaths	Pop.	Deaths	Pop.	Deaths
40-44	656	18	145	2	4531	149	3410	124
45-49	359	22	104	4	3030	169	2239	140
50-54	249	19	98	3	2267	193	1851	187
55-59	632	55	372	38	4682	576	3270	514
60-64	1067	117	846	113	6052	1001	3791	778
65-69	897	170	949	173	3880	901	2421	689
70-74	668	179	824	212	2033	613	1195	432
75-80	361	120	667	243	871	337	436	214
> 80	274	120	537	253	345	189	113	63

Our model is as follows. There are, in total, 36 age/smoking groups. Let the number of deaths in age/smoking group i be y_i from a population of n_i. Then we suppose that, given p_i, y_i is an observation on a binomial random variable

$$Y_i \sim \text{Bin}(n_i, p_i)$$

independently for all i. Then we suppose that

$$\log_e \left(\frac{p_i}{1-p_i} \right) = \alpha_{g(i)} + \beta_{g(i)}(x_i - 60)$$

where $g(i)$ is the smoking group to which age/smoking group i belongs with $g = 1$ for "Nonsmokers", $g = 2$ for "Cigar and pipe", $g = 3$ for "Cigarette and other" and $g = 4$ for

"Cigarette only." The covariate x_i is the lower limit of the age range in years for the age group to which age/smoking group i belongs. So each x_i is one of 40, 45, ..., 75, 80.

Our prior distribution is as follows. Given the value of α_0, we have

$$\alpha_g \sim N(\alpha_0, 5)$$

with $\alpha_1, \ldots, \alpha_4$ conditionally independent. Given the value of β_0, we have

$$\beta_g \sim N(\beta_0, 0.1)$$

with $\beta_1, \ldots, \beta_4$ conditionally independent. Finally we have independent normal priors for α_0 and β_0 :

$$\begin{aligned} \alpha_0 &\sim N(-2, 100), \\ \beta_0 &\sim N(0, 0.125). \end{aligned}$$

A suitable BUGS model file is listed below. This file may be downloaded from the Module Web Page at

`http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/healthbug.txt`

```
model smoking

{ for (i in 1:36)
      {deaths[i]~dbin(p[i],n[i])
       logit(p[i])<-alpha[group[i]]+beta[group[i]]*(age[i]-60)
       }

 for (g in 1:4)
     {alpha[g]~dnorm(alpha0,0.2)
      beta[g]~dnorm(beta0,10)
      }

 alpha0~dnorm(-2,0.01)
 beta0~dnorm(0.0,8)
 }
```

A suitable data file may be downloaded from the Module Web Page at

`http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/healthdata.txt`

(a) Write down the prior mean and variance for $\log_e(p_i/(1-p_i)$ where p_i is the probability of death for a nonsmoker in the 55-59 age group.

Solution

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \alpha_1 - 5\beta_1$$

Now $\mathrm{E}(\alpha_1) = -2$ and $\mathrm{E}(\beta_1) = 0$ so the prior mean is

$$\mathrm{E}(\eta_i) = -2.$$

The variances are

$$\text{var}(\alpha_1) = \frac{1}{0.01} + \frac{1}{0.2} = 100 + 5 = 105$$

and

$$\text{var}(\beta_1) = \frac{1}{8} + \frac{1}{10} = 0.125 + 0.1 = 0.225.$$

So the required prior variance is

$$\text{var}(\eta_i) = 105 + 25 \times 0.225 = 110.625.$$

(2 marks)

(b) Find a prior 95% interval for p_i where p_i is the probability of death for a nonsmoker in the 55-59 age group.

Solution

95% interval for η_i is $-2 \pm 1.96\sqrt{110.625}$. That is $-22.61497 < \eta_i < 18.61497$.

Now

$$p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

so 95% interval for p_i is

$$1.5 \times 10^{-10} < p_i < 1 - 1.5 \times 10^{-10}.$$

(In effect $0 < p_i < 1$).

(2 marks)

(c) Use BRugs to evaluate the posterior distribution. *Read the rest of this question to see what you need to do before you run the program.*

i. Write down the commands which you use.

ii. Find the posterior mean of β_0.

iii. Find a 95% posterior interval for α_0.

iv. Find a 95% posterior interval for $\alpha_3 - \alpha_1$ and comment briefly on what this tells us about the risk involved in cigarette smoking. *(To do this part you may wish to modify the model file slightly).*

Solution

First I edited the model file to add the line

```
diff<-alpha[3]-alpha[1]
```

in order to answer part iv. The new file is called `healthbuga.txt`.

Here is a complete listing of the session (although it is only necessary for a candidate to give the commands). First I checked convergence. Then I did a second run to compute the posterior.

```
> modelCheck("healthbuga.txt")
model is syntactically correct
> modelData("healthdata.txt")
data loaded
> modelCompile(2)
model compiled
```

```
> modelGenInits()
initial values generated, model initialized
> samplesSet(c("alpha0","beta0","alpha","beta"))
monitor set for variable 'alpha0'
monitor set for variable 'beta0'
monitor set for variable 'alpha'
monitor set for variable 'beta'
> modelUpdate(1000)
1000 updates took 0 s
> samplesHistory("alpha0")
Waiting to confirm page change...
> samplesHistory("beta0")
Waiting to confirm page change...
> samplesHistory("alpha")
Waiting to confirm page change...
Waiting to confirm page change...
> samplesHistory("beta")
Waiting to confirm page change...
Waiting to confirm page change...
> modelCheck("healthbuga.txt")
model is syntactically correct
> modelData("healthdata.txt")
data loaded
> modelCompile(2)
model compiled
> modelGenInits()
initial values generated, model initialized
> modelUpdate(500)
500 updates took 0 s
> samplesSet(c("alpha0","beta0","diff"))
monitor set for variable 'alpha0'
monitor set for variable 'beta0'
monitor set for variable 'diff'
> modelUpdate(10000)
10000 updates took 0 s
> samplesStats("alpha0")
         mean    sd MC_error val2.5pc median val97.5pc start sample
alpha0 -1.718 1.109 0.007984   -3.894  -1.72    0.4505   501  20000
> samplesStats("beta0")
        mean     sd  MC_error val2.5pc  median val97.5pc start sample
beta0 0.0707 0.1444 0.0009801  -0.2116 0.06996    0.3535   501  20000
> samplesStats("diff")
       mean      sd  MC_error val2.5pc median val97.5pc start sample
diff 0.2876 0.04976 0.0004068     0.19 0.2876     0.384   501  20000
```

So, the posterior mean of β_0 is 0.0707.

The 95% posterior interval for α_0 is $-3.894 < \alpha_0 < 0.4505$.

The 95% posterior interval for $\alpha_3 - \alpha_1$ is $0.190 < \alpha_3 - \alpha_1 < 0.384$. So, for men in the 60-64 age-group, we can conclude that there is almost certainly an increased risk of death in the "Cigarette and Other" group compared to the nonsmokers. (This is actually the log odds ratio so the 95% interval for the odds ratio would be $1.21 < R < 1.47$, *ie* a 21%-47% increase in the odds in favour of death).

(10 marks)

2. The data below come from an ecological study. Visits were made to four sites on two occasions each and the numbers of individuals observed of each of three species were counted.

Species	Site	Count	Species	Site	Count
1	1	3	1	1	3
2	1	1	2	1	0
3	1	6	3	1	6
1	2	10	1	2	8
2	2	6	2	2	1
3	2	21	3	2	21
1	3	0	1	3	0
2	3	0	2	3	1
3	3	3	3	3	6
1	4	5	1	4	5
2	4	2	2	4	0
3	4	12	3	4	13

Our model is as follows. The counts, *ie* the numbers of individuals observed, are $y_1, \ldots, y_{24}$, where y_i is regarded as an observation from the Poisson distribution

$$Y_i \sim \text{Po}(\lambda_i)$$

where

$$\log_e(\lambda_i) = \mu + \alpha_{a(i)} + \beta_{b(i)}$$

and $a(i)$ denotes the site (1,2,3,4) at which observation i was made and $b(i)$ denotes the species (1,2,3) for observation i.

We regard $\alpha_1, \ldots, \alpha_4$ as random effects with

$$\alpha_j \sim N(0, \tau_a^{-1})$$

independently, given τ_a.

We regard $\beta_1, \ldots, \beta_3$ as random effects with

$$\beta_j \sim N(0, \tau_b^{-1})$$

independently, given τ_b.

We have independent prior distributions for μ, τ_a and τ_b with a normal prior for μ :

$$\mu \sim N(0, 5)$$

and gamma priors for τ_a and τ_b :

$$\tau_a \sim \text{Ga}(1, 0.1),$$
$$\tau_b \sim \text{Ga}(1, 0.1).$$

The data are available in a file ready for use in BRugs. The file may be downloaded from

`http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/ecodata.txt`

The file contains three columns. Its contents are reproduced below.

```
species[] site[] count[]
       1     1      3
       2     1      1
       3     1      6
       1     2     10
       2     2      6
       3     2     21
       1     3      0
       2     3      0
       3     3      3
       1     4      5
       2     4      2
       3     4     12
       1     1      3
       2     1      0
       3     1      6
       1     2      8
       2     2      1
       3     2     21
       1     3      0
       2     3      1
       3     3      6
       1     4      5
       2     4      0
       3     4     13
END
```

Use BRugs to evaluate the posterior distribution. *Read the rest of this question to see what you need to do before you run the program.*

(a) Write down your BRugs model specification.
Possible solution

```
model eco

{ for (i in 1:24)
      {count[i]~dpois(lambda[i])
       log(lambda[i])<-mu+alpha[site[i]]+beta[species[i]]
        }

  for (s in 1:4)
      {alpha[s]~dnorm(0,tau.site)
       }

  for (s in 1:3)
      {beta[s]~dnorm(0,tau.species)
       }

  mu~dnorm(0,0.2)

  tau.site~dgamma(1,0.1)
  tau.species~dgamma(1,0.1)

 }
```

(3 marks)

(b) Write down the commands which you use.

Solution

Here is a complete listing of the session (although it is only necessary for a candidate to give the commands). First I checked convergence. Then I did a second run to compute the posterior.

```
> modelCheck("ecobug.txt")
model is syntactically correct
> modelData("ecodata.txt")
data loaded
> modelCompile(2)
model compiled
> modelGenInits()
initial values generated, model initialized
> samplesSet(c("alpha","beta","mu","tau.site","tau.species"))
monitor set for variable 'alpha'
monitor set for variable 'beta'
monitor set for variable 'mu'
monitor set for variable 'tau.site'
monitor set for variable 'tau.species'
> modelUpdate(1000)
1000 updates took 0 s
> samplesHistory("alpha")
Waiting to confirm page change...
Waiting to confirm page change...
> samplesHistory("beta")
Waiting to confirm page change...
> samplesHistory("mu")
Waiting to confirm page change...
> samplesHistory("tau.site")
Waiting to confirm page change...
> samplesHistory("tau.species")
Waiting to confirm page change...
> modelCheck("ecobug.txt")
model is syntactically correct
> modelData("ecodata.txt")
data loaded
> modelCompile(2)
model compiled
> modelGenInits()
initial values generated, model initialized
> modelUpdate(10000)
10000 updates took 1 s
> samplesSet(c("tau.site","mu"))
monitor set for variable 'tau.site'
monitor set for variable 'mu'
> modelUpdate(30000)
30000 updates took 5 s
> samplesHistory("mu")
Waiting to confirm page change...
> samplesHistory("tau.site")
Waiting to confirm page change...
> samplesStats("mu")
```

```
    mean     sd MC_error val2.5pc median val97.5pc start sample
mu 1.147 0.6297  0.02272  -0.2152  1.162     2.393 10001  60000
> samplesStats("tau.site")
          mean    sd MC_error val2.5pc median val97.5pc start sample
tau.site 2.825 2.109  0.01771   0.4216  2.307     8.246 10001  60000
```

(2 marks)

(c) Find the posterior mean of τ_a.

Solution

The posterior mean for τ_a is 2.825.

(2 marks)

(d) Find a 95% posterior interval for μ.

Solution

The 95% posterior interval for μ is $-0.2152 < \mu < 2.393$.

(2 marks)

(e) Comment on the behaviour of the MCMC sampler (e.g. convergence, mixing).

Solution

Mixing is poor. Convergence is especially poor for μ. For this reason I used a long burn-in and a large number of samples.

(2 marks)

References

Best, E.W.R. and Walker, C.B., 1964. A Canadian study of smoking and health. *Canadian Journal of Public Health*, **55**, 1.