MAS8303 Modern Bayesian Inference

Outline Solution to Exercise 3.5

Table 1 shows the numbers of patients undergoing surgery and the numbers who died in the hospital following surgery in two areas of the USA, broken down by age-group and sex. The data are taken from Mosteller and Tukey (1977).

We propose the following model. There are four area-sex groups:

Group 1 : Males in Area 1.

Group 2 : Females in Area 1.

Group 3 : Males in Area 2.

Group 4 : Females in Area 2.

Given the model parameters, the number of deaths in Area-Sex Group j and Age-group k has a binomial($n_{j,k},\ p_{j,k}$) distribution where $n_{j,k}$ is the number of patients undergoing surgery and

$$\log\left(\frac{p_{j,k}}{1-p_{j,k}}\right) = \alpha_j + \beta_j(x_k - 50)$$

where x_k is the mid-point of the age-range for age-group k.

We need to make inferences about the eight model parameters, $\alpha_1, \ldots, \alpha_4, \beta_1, \ldots, \beta_4$.

1. Suppose that we consider "typical" patients aged 50. Suppose that for such patients, the probability p_0 of death is α_0 and we give α_0 a normal prior distribution. Suppose that, in our prior beliefs, $\Pr(p_0 < 0.02) = \Pr(p_0 > 0.10) = 0.025$.

   For a "typical" patient aged 50 we have

   $$\log\left(\frac{p_0}{1-p_0}\right) = \alpha_0 + \beta_0(50-50) = \alpha_0.$$

   We have a symmetric 95% interval for p_0 which is

   $$0.02 < p_0 < 0.1.$$

   Hence a symmetric 95% interval for α_0 is

   $$\log\left(\frac{0.02}{0.98}\right) < \alpha_0 < \log\left(\frac{0.1}{0.9}\right).$$

   That is

   $$-3.892 < \alpha_0 < -2.197.$$

   Hence our prior mean for α_0 is $-(3.892 + 2.197)/2 = \underline{-3.0445}$ and our prior variance is

   $$\left(\frac{3.892 - 2.197}{2 \times 1.96}\right)^2 = \underline{0.1870}.$$

2. Our joint prior distribution for $\alpha_1, \ldots, \alpha_4$ can be represented as follows. We write

   $$\begin{array}{rcl} \alpha_j \mid \bar{\alpha} & \sim & N(\bar{\alpha},\ V_{\alpha,1}) \quad \text{for } j = 1, \ldots, 4. \\ \bar{\alpha} & \sim & N(m_\alpha,\ V_{\alpha,0}). \end{array}$$

   Here $\alpha_1, \ldots, \alpha_4$ are conditionally independent given $\bar{\alpha}$. We choose to make $V_{\alpha,0} = V_{\alpha,1}$ and $V_{\alpha,0} + V_{\alpha,1}$ gives the prior variance of α_0.

   Since we have $V_{\alpha,0} + V_{\alpha,1} = \text{var}(\alpha_0) = 0.1870$ and $V_{\alpha,0} = V_{\alpha,1}$, we must have $V_{\alpha,0} = V_{\alpha,1} = \underline{0.0935}$.

3. We propose a matching structure for $\beta_1, \ldots, \beta_4$ with $\beta_1, \ldots, \beta_4$ independent of $\alpha_1, \ldots, \alpha_4$ in the prior.

   $$\begin{array}{rcl} \beta_j \mid \bar{\beta} & \sim & N(\bar{\beta},\ V_{\beta,1}) \quad \text{for } j = 1, \ldots, 4. \\ \bar{\beta} & \sim & N(m_\beta,\ V_{\alpha,0}). \end{array}$$

   Here $\beta_1, \ldots, \beta_4$ are conditionally independent given $\bar{\beta}$. We choose to make $V_{\beta,0} = V_{\beta,1}$ and $V_{\beta,0} + V_{\beta,1} = 0.0004$.

   Since we have $V_{\beta,0} + V_{\beta,1} = 0.0004$ and $V_{\beta,0} = V_{\beta,1}$, we must have $V_{\beta,0} = V_{\beta,1} = \underline{0.0002}$.

   The value of m_β is 0.0.

4. A suitable BRugs model specification is shown in Figure 1. Note that, for the precisions, $1/0.0935 = 10.7$ and $1/0.0002 = 5000$.

5. The data are available in a file called `surgicaldata.txt`. The data have been arranged into four columns as follows.

   - `group`: the area-sex group number as above.
   - `age`: the midpoint of the age range for the age-group.
   - `patients`: the number of patients undergoing surgery.
   - `deaths`: the number of deaths.

   Use BRugs to find the posterior distribution of the model parameters. Check convergence of the sampler.

   An initial run of BRugs was used to check the convergence behaviour as follows.

```
> modelCheck("surgicalbug.txt")
> modelData("surgicaldata.txt")
> modelCompile(2)
> modelGenInits()
> samplesSet(c("alphabar","betabar","alpha","beta"))
> modelUpdate(2000)
> samplesHistory("alphabar")
> samplesHistory("betabar")
> samplesHistory("alpha")
> samplesHistory("beta")
```

Area 1

	Total undergoing surgery		Number dying	
Age	Males	Females	Males	Females
5-14	4272	3911	9	11
15-24	2835	2989	23	5
25-34	2785	2606	19	8
35-44	1930	1886	16	15
45-54	1497	1524	59	40
55-64	960	1013	101	52
65-75	652	855	185	118
76-83	186	287	97	108

Area 2

	Total undergoing surgery		Number dying	
Age	Males	Females	Males	Females
5-14	1739	1758	5	2
15-24	1233	1244	14	1
25-34	989	1004	8	3
35-44	897	922	9	13
45-54	921	961	28	15
55-64	686	739	68	37
65-75	611	784	159	73
76-83	189	290	86	88

Table 1: Deaths following surgery in two areas of the USA

The graphs showed that convergence was rapid and that mixing was reasonably good. (The graphs are omitted). Therefore a burn-in of 1000 iterations would be sufficient and 10000 further samples should provide a reasonable approximation to the posterior distribution. (There are further checks of these things available but we have not covered them).

6. Posterior summaries were calculated as follows.

```
> modelCheck("surgicalbug.txt")
> modelData("surgicaldata.txt")
> modelCompile()
> modelGenInits()
> modelUpdate(1000)
> samplesSet(c("alphabar","betabar","alpha","beta"))
> modelUpdate(10000)
> samplesStats(c("alphabar","betabar","alpha","beta"))
```

The results are presented in Table 2.

The "Monte Carlo Error" gives an indication of the likely accuracy of the numerical approximation obtained using the Gibbs sampler. This could be reduced by increasing the number of iterations.

7. To find the posterior distribution of $\log(p_1^*/p_3^*)$ where p_1^* is the probability of death for a fifty-year-old male in area 1 and p_3^* is the probability of death for a fifty-year-old male in area 2, one way is simply to add the following lines

```
 for (k in 1:4)
   {pstar[k]<-exp(alpha[k])/(1+exp(alpha[k]))
    }
```

```
model surgical

{
 for (i in 1:32)
     {deaths[i]~dbin(p[i],patients[i])
      logit(p[i])<-alpha[group[i]]+beta[group[i]]*(age[i]-50)
      }

 for (j in 1:4)
     {alpha[j]~dnorm(alphabar,10.7)
      beta[j]~dnorm(betabar,5000)
      }

 alphabar~dnorm(-3.0445,10.7)
 betabar~dnorm(0.0,5000)


  }
```

Figure 1: BRugs model specification.

Quantity	Mean	Std.dev.	Monte Carlo Error	Median	95% Interval Lower	Upper
$\bar{\alpha}$	-3.30200	0.140700	0.00489	-3.30200	-3.58400	-3.02300
$\bar{\beta}$	0.07341	0.006413	0.00008	0.07346	0.06069	0.08598
α_1	-3.01000	0.057340	0.00072	-3.00800	-3.12700	-2.89800
α_2	-3.60600	0.071930	0.00090	-3.60500	-3.74700	-3.46600
α_3	-3.01800	0.076160	0.00091	-3.01800	-3.16900	-2.87200
α_4	-3.81600	0.103000	0.00146	-3.81400	-4.02000	-3.62000
β_1	0.09599	0.003170	0.00004	0.09595	0.08970	0.10230
β_2	0.09183	0.003596	0.00004	0.09183	0.08485	0.09883
β_3	0.09031	0.003985	0.00005	0.09027	0.08267	0.09820
β_4	0.08875	0.004838	0.00007	0.08863	0.07952	0.09855

Table 2: Posterior summaries.

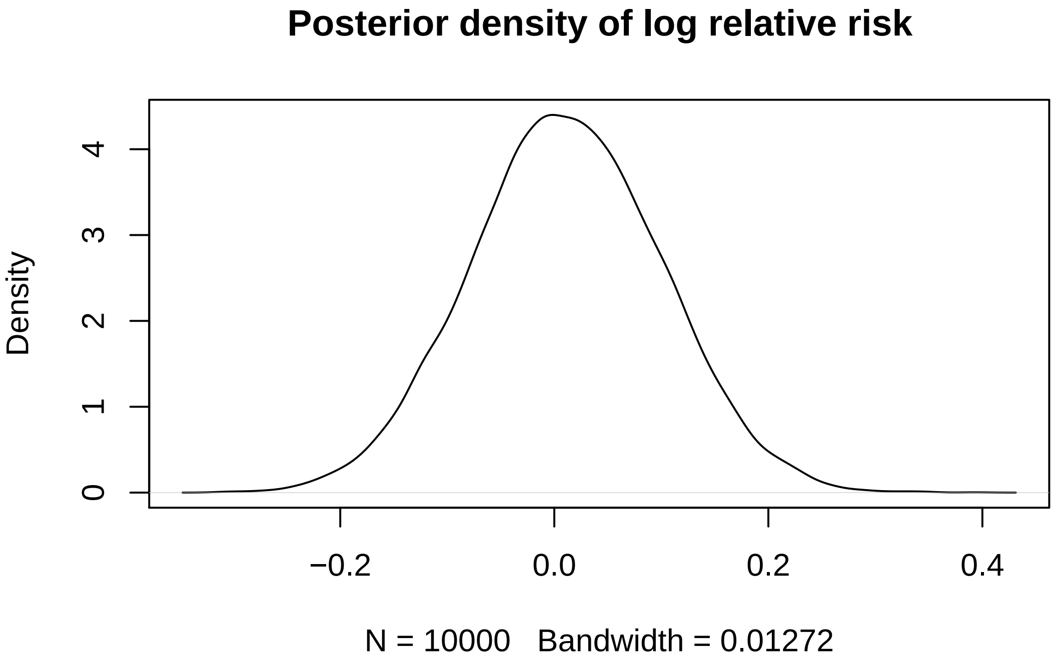

Figure 2: Posterior density of $\log(p_1^*/p_3^*)$.

```
 z<-log(pstar[1]/pstar[3])
```

to the BRugs model specification and monitor this quantity. (Alternatively we could also extract samples and do the necessary calculations in R rather than running the sampler again). The calculations were done as follows.

```
> modelCheck("surgicalbug.txt")
> modelData("surgicaldata.txt")
> modelCompile()
> modelGenInits()
> modelUpdate(1000)
> samplesSet("z")
> modelUpdate(10000)
> samplesStats("z")
```

The posterior mean was found to be 0.01124 and the posterior standard deviation was found to be 0.08921. The posterior density is shown in Figure 2. We can see that the posterior distribution is centred close to zero so there is no strong suggestion that the risk for a 50-year-old male patient is greater in Area 1 or in Area 2. In fact 95% of the posterior probability is between about -0.135 and 0.185. This range corresponds to a range of p_1^*/p_3^* (i.e. a *relative risk*) of between 0.87 and 1.20.

The graph was produced as follows.

```
> z<-samplesSample("z")
> pdf("relrisk.pdf",height=4)
> plot(density(z),main="Posterior density of log relative risk")
> dev.off()
```