

Bayesian Statistics: Assessment Exercise 4

Marks for this assignment: 11 of course total of 88 (12.5%)

See Moodle for submission and deadline

For this exercise, you need to submit a single text file or Word document that encompasses the answers to all of the questions below, with the section number (a) to (e) as a header for each part. Where JAGS code is asked for, you do NOT need to specify the R code needed to run it.

You have been given some data representing counts (Count) of aphids on 100 (N) crop plants in 20 (F) fields (Field). Some of the plants were treated with insecticide and some were not (Treatment). You also know how many days each plant has been exposed to potential sources of infection (Time). The data is as follows:

```
Count <- c(6, 14, 1, 5, 11, 1, 3, 2, 1, 1, 3, 3, 2, 8, 5, 10, 0, 3, 3, 6,
1, 6, 1, 19, 8, 10, 9, 6, 3, 2, 4, 2, 3, 16, 11, 3, 8, 2, 4, 7, 4, 2, 8,
4, 1, 2, 5, 6, 2, 10, 4, 5, 9, 7, 10, 13, 8, 2, 9, 6, 6, 4, 3, 0, 0, 2,
0, 5, 2, 6, 3, 1, 11, 3, 4, 5, 3, 2, 2, 5, 4, 1, 3, 10, 7, 0, 2, 2, 11,
7, 1, 11, 1, 3, 3, 4, 6, 6, 1, 10);
```

```
Field <- c(4, 5, 17, 10, 20, 7, 6, 13, 16, 8, 19, 8, 2, 6, 4, 16, 17, 9,
1, 17, 5, 6, 19, 11, 17, 5, 15, 11, 14, 16, 16, 7, 20, 11, 7, 13, 8, 4,
3, 1, 19, 7, 6, 3, 8, 9, 16, 10, 10, 10, 19, 10, 15, 5, 3, 3, 9, 7, 1, 5,
6, 10, 9, 14, 9, 7, 19, 1, 19, 12, 1, 4, 20, 20, 6, 18, 20, 17, 2, 8, 5,
10, 10, 1, 10, 2, 8, 14, 12, 18, 8, 8, 6, 16, 14, 15, 3, 3, 19, 3);
```

```
Treatment <- c(2, 2, 1, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 1,
2, 1, 2, 1, 2, 2, 2, 1, 1, 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2,
1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 1, 1, 1, 1,
1, 1, 1, 2, 1, 2, 2, 2, 2, 1, 1, 1, 2, 2, 1, 2, 2, 2, 1, 1, 2, 2, 2, 1,
2, 1, 2, 1, 1, 2, 1, 1, 2);
```

```
Time <- c(1.86, 2.55, 1.55, 2.35, 2.76, 2.04, 2.11, 1.81, 1.88, 1.09,
2.36, 1.67, 1.01, 2.64, 2.14, 1.98, 1.38, 2.46, 2.68, 2.05, 1.42, 3,
2.07, 2.59, 2.61, 1.79, 2.43, 1.56, 1.84, 2.13, 2.41, 2, 1.57, 1.56,
2.72, 2.39, 2.36, 1.58, 2.14, 2.78, 2.54, 2.85, 1.62, 2.12, 2.66, 2.13,
1.6, 2.14, 1.7, 2.13, 2.17, 2.81, 1.25, 1.99, 2.78, 2.64, 2.87, 1.15,
1.15, 2.78, 1.01, 1.26, 1.12, 1.4, 1.53, 1.12, 1.08, 1.07, 1.57, 1.91,
1.28, 1.92, 2.78, 1.34, 2.11, 1.23, 2.42, 1.98, 2.96, 1.84, 1.78, 2.92,
1.4, 2.51, 1.41, 2.54, 2.44, 1, 1.26, 1.53, 1.63, 2.57, 1.16, 1.3, 2.91,
2.34, 1.62, 2.75, 2.4, 2.34);
```

```
N <- 100;
```

```
F <- 20;
```

This full model can be used to describe the data:

```
model{  
  for(i in 1:N){  
    Count[i] ~ dpois(lambda[i])  
    log(lambda[i]) <- intercept + treatmenteffect[Treatment[i]] +  
    exposureeffect * Time[i] + fieldeffect[Field[i]]  
  }  
  for(f in 1:F){  
    fieldeffect[f] ~ dnorm(0, tau)  
  }  
  intercept ~ dnorm(0, 10^-6)  
  treatmenteffect[1] <- 0  
  treatmenteffect[2] ~ dnorm(0, 10^-6)  
  exposureeffect ~ dnorm(0, 10^-6)  
  tau ~ dgamma(0.001, 0.001)  
}
```

(a) Which of the model parameters shown would you specify initial values for? Give suitable initial values for each of these parameters for two chains, using valid R code.

[Hint: you can use the model and data provided to check your initial values]

(3 marks)

(b) The biologist is worried about model parsimony, and that he may have collected information on one or more predictor variables that are not related to his outcome variable at all. Adapt the JAGS code given to represent the simplest model that you would use for this data, and give your model below.

(3 marks)

[Continued on next page]

Examine the output from the four nested models (unrelated to the data in question a/b) given below:

```
> model1          # The full model
```

JAGS model summary statistics from 20000 samples (chains = 2; burnin = 5000):

	Lower95	Median	Upper95	Mean	SD	MCerr	MC%ofSD	SSEff	AC.10	psrf
breedeffect[1]	0	0	0	0	0					
breedeffect[2]	-0.31313	0.11168	0.54288	0.11182	0.21924	0.0046788	2.1	2196	0.14017	1.0005
breedeffect[3]	-0.41892	0.0014235	0.39952	0.0020591	0.20993	0.0044612	2.1	2214	0.13221	1
intercept	0.62796	0.99239	1.3377	0.99146	0.18164	0.0043813	2.4	1719	0.18546	1.0002
sexeffect[1]	0	0	0	0	0					
sexeffect[2]	0.99756	1.2279	1.4753	1.031	0.17286	0.0024157	1.4	5120	0.020589	1.0002

Model fit assessment (DIC): 317.8779

Total time taken: 7 seconds

```
> model2          # An intermediate model
```

JAGS model summary statistics from 20000 samples (chains = 2; burnin = 5000):

	Lower95	Median	Upper95	Mean	SD	MCerr	MC%ofSD	SSEff	AC.10	psrf
breedeffect[1]	0	0	0	0	0					
breedeffect[2]	-0.42896	0.0075009	0.40141	0.0062185	0.21368	0.0043305	2	2435	0.11813	1.0012
breedeffect[3]	-0.45256	-0.042739	0.37098	-0.044619	0.20902	0.0043197	2.1	2341	0.13335	1.0004
intercept	1.2234	1.534	1.8632	1.5377	0.16261	0.003704	2.3	1927	0.15807	1.0004
sexeffect[1]	0	0	0	0	0					
sexeffect[2]	0	0	0	0	0					

Model fit assessment (DIC): 333.3096

Total time taken: 5 seconds

```
> model3          # An intermediate model
```

JAGS model summary statistics from 20000 samples (chains = 2; burnin = 5000):

	Lower95	Median	Upper95	Mean	SD	MCerr	MC%ofSD	SSEff	AC.10	psrf
breedeffect[1]	0	0	0	0	0					
breedeffect[2]	0	0	0	0	0					
breedeffect[3]	0	0	0	0	0					
intercept	0.81661	1.0284	1.2389	1.0296	0.10802	0.0014465	1.3	5576	0.013488	1.0001
sexeffect[1]	0	0	0	0	0					
sexeffect[2]	0.60362	0.8041	1.0647	1.025	0.1695	0.0023549	1.4	5181	0.013465	1.0002

Model fit assessment (DIC): 314.2675

Total time taken: 5 seconds

```
> model4          # The simplest model
```

JAGS model summary statistics from 20000 samples (chains = 2; burnin = 5000):

	Lower95	Median	Upper95	Mean	SD	MCerr	MC%ofSD	SSEff	AC.10	psrf
breedeffect[1]	0	0	0	0	0					
breedeffect[2]	0	0	0	0	0					
breedeffect[3]	0	0	0	0	0					
intercept	1.355	1.518	1.6803	1.5186	0.082593	0.00073768	0.9	12536	-0.012968	0.99996
sexeffect[1]	0	0	0	0	0					
sexeffect[2]	0	0	0	0	0					

Model fit assessment (DIC): 349.4509

Total time taken: 3 seconds

(c) Which part of these outputs best indicates the empirical fit of these models to the data? Describe in your own words what this statistic means, and what two pieces of information are used to calculate it.

(2 marks)

(d) Based on your answer to (c), which model (1 to 4) is preferred?

(1 mark)

(e) Discuss any caveats to your response for (d), and the effect that this may have on the posterior inference made from these data (refer to the four sets of model output given above where appropriate).

(2 marks)