

UNIVERSITY OF WOLLONGONG
School of Mathematics and Applied Statistics
STAT902. Advanced Data Analysis

ASSIGNMENT 8

Due: 5:00pm Monday 10th May, 2010.

Local students: please put under lecturer's door.

Remote students: please e-mail or post by this date.

BUGS preliminaries

This assignment deals mainly with fitting and making inference for Bayesian statistical models using a version of the **BUGS** (**B**ayesian **U**sing **G**ibbs **S**ampling) suite of software. In particular it shows how to do BUGS-based analyses inside **R** using the **WinBUGS** package (which only runs under the **Windows** operating system) and the **R** package **BRugs**. The computers in the Horner Laboratory now have both **WinBUGS** and **BRugs** installed. If you wish to install these on your own computer then please see the lecturer for advice about this. This assignment assumes that you are using the **Windows** operating system on a computer that has **WinBUGS** and **BRugs** installed.

1. Obtain the following files from the **Computer Code and Data** page on the course website: `binomBeta.Rs`, `binomBetaChk.Rs`, `binomBetaModel.txt`, `bpd.txt`, `bpdBayes.Rs`, `bpdBayesModel.txt` and `summariseMCMC.r`.

Files with names of the form `*.Rs` are **R** scripts for running BUGS-based analyses.

Files with names of the form `*Model.txt` are support files containing the model specification using the BUGS language syntax. While BUGS code has some similarities with **R**, it also has some differences. Note that each of the `*.Rs` files calls a particular `*Model.txt` file.

The file `summariseMCMC.r` contains a function named `summariseMCMC()` to facilitate the summary of Markov Chain Monte Carlo (MCMC) output from **BRugs**.

The file `bpd.txt` contains the data on bronchopulmonary dysplasia (BPD) and birth-weights used to introduce logistic regression earlier in the course.

2. Start an **R** session and make sure that the working directory corresponds to that where the abovementioned files are located.
3. Enter the command `source("binomBeta.Rs")`. This should facilitate Bayesian inference for the binomial parameter p of a binomial data set ($n = 20$ Bernoulli observations) with a uniform ($\text{Beta}(1,1)$) prior placed on p . The script produces a 4-panel summary of the MCMC output (using `summariseMCMC()`). Look at the code in the files `binomBetaModel.txt` and `binomBeta.Rs` to see how **BUGS** and **BRugs** work.
4. We don't really need **BUGS** and MCMC to perform Bayesian inference for the binomial model with Beta prior. It is a relatively rare example of a Bayesian model with an analytic posterior distribution. For the y data in `binomBeta.Rs` it is a simple exercise to show that the posterior is the $\text{Beta}(4,18)$ distribution. Enter the command

`source("binomBetaChk.Rs")` to check that the MCMC samples are, indeed, coming from the correct distribution.

5. Enter the command `source("bpdBayes.Rs")`. This fits the Bayesian logistic regression model with diffuse normal priors:

$$\text{logit}\{P(\text{BPD}_i = 1)\} = \beta_0 + \beta_1 \text{sbirthweight}_i, \quad 1 \leq i \leq 223$$

$$\beta_0, \beta_1 \stackrel{\text{ind.}}{\sim} N(0, 10^8)$$

Note that `sbirthweighti` is the *standardised* birthweight data — transformed to have mean zero and variance 1. This is recommended for continuous in Bayesian regression modelling with diffuse priors to make the analysis scale invariant. With a little more coding the results could be presented in terms of the original birthweight variable, but we will leave this out for now.

6. **WARNING I:** Note that the `dnorm` function in BUGS for specifying that a parameter has a normal distribution uses *precision* parameterisation rather than the more common *variance* parameterisation. The precision is the reciprocal of the variance. For example, to specify $\theta \sim N(0, 40)$ in BUGS the command `theta ~ dnorm(0, 0.025)` should be used.
7. **WARNING II:** BRugs is a very new software product – only a few years old. It does not yet give very specific errors messages. This means that finding mistakes in your BUGS and BRugs code is challenging. It is important to check that all variables appearing in BUGS and referenced in BRugs with precisely the same name. All parameters should be initialised. If you accidentally type `beta0` as `beta0` then the programme won't run (and won't tell you why!).
8. **WARNING III:** Markov chain Monte Carlo (MCMC) is still in its infancy as a statistical tool. While it provides satisfactory solutions to difficult inferential problems a lot of the time, it is prone to breakdown and misleading results from time to time. Diagnostic checks (such as those produced by `summariseMCMC()`) are recommended. Exactly what and how much diagnosis on the MCMC output should be done is a topic of ongoing discussion and research in the Bayesian analysis and MCMC literature. In STAT902 we will stay with the rudimentary diagnostics of `summariseMCMC()`.

'Hand-in' part of the assignment

1. Alana Connelly is the editor for the *Unicorn* book series and is interested in

$$\lambda = \text{mean number of typographical errors per page}$$

in the books. Over the next several weeks she plans to sample 25 pages from recently published Unicorn books and record the number of typographical errors. Let $\mathbf{y} = (y_1, \dots, y_{25})$ be the vector of typographical error counts. Alana postulates the following Bayesian model for the data:

$$[\mathbf{y}|\lambda] = \prod_{i=1}^{25} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \quad (\text{i.e. } y_i|\lambda \stackrel{\text{ind.}}{\sim} \text{Poisson}(\lambda))$$

and imposes the prior

$$[\lambda] = \frac{1}{16} \lambda^2 e^{-\lambda/2}, \quad \lambda > 0.$$

Note that the prior density function of λ is the Gamma(3,2) distribution (in the notation of the *Likelihood Theory and Methods* notes).

IMPORTANT: Before attempting this question you need to be aware of the fact that textbooks and software packages differ in their parametrisations of the Gamma distribution. The *Likelihood Theory and Methods* notes and other course material sticks with the parameterisation:

$$[x; \alpha, \beta] = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha) \beta^\alpha}, \quad x > 0.$$

In this parametrisation β plays the role of a *scale* parameter. An alternative parametrisation has the density function being (for parameters $A, B > 0$):

$$[x; A, B] = \frac{B^A e^{-Bx} x^{A-1}}{\Gamma(A)}, \quad x > 0.$$

Here B is usually called a *rate* parameter. Comparing the two parametrisations we see that the *shape* parameters α and A are the same, but the scale and rate parameters have a reciprocal relationship: $\beta = 1/B$. In R, typing `help(dgamma)` or `help(rgamma)` reveals that both parametrisations are supported. However, in BUGS the rate parametrisation is used. All of this needs to be taken into account for correct completion of this assignment question.

- (a) Find the posterior density function of λ in terms of \mathbf{y} .
- (b) The observed data are:

$$\mathbf{y} = (7, 6, 7, 5, 7, 3, 4, 10, 9, 8, 10, 3, 7, 5, 6, 13, 10, 9, 9, 4, 12, 2, 7, 7, 3)$$

Obtain the posterior density function for these data. Using R make a plot of the prior and posterior densities on the same axes. Use the `lwd` parameter to distinguish the curves. The following code illustrates this for a simple line and parabola:

```
xg <- seq(-1,1,length=101)
y1g <- xg ; y2g <- xg^2
plot(xg,y1g,type="l") ; lines(xg,y2g,lwd=3)
```

- (c) What is the posterior mean $E(\lambda|\mathbf{y})$ both for general \mathbf{y} and for the observed data?
- (d) Use BRugs and WinBUGS to obtain 1000 draws from the posterior distribution. Based on this sample approximate the posterior mean and the posterior density. Make a plot for comparison of the approximate posterior density and the exact posterior density.

Hint: Modify code in the `binomBeta*` files used in the BUGS preliminaries.

2. (a) Consider the Bayesian logistic regression model

$$\text{logit}\{P(y_i = 1)\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i}. \quad 1 \leq i \leq n$$

where $\beta_0, \dots, \beta_7 \stackrel{\text{i.i.d.}}{\sim} N(0, 10^8)$. Assume the predictors x_{ji} are fixed (non-random). Find an expression for the posterior distribution of β_5 :

$$[\beta_5 | y_1, \dots, y_n] = [\beta_5 | \mathbf{y}]$$

where $\mathbf{y} = (y_1, \dots, y_n)$. It will not be possible to obtain a closed form expression without integrals, but obtain as simple expression as possible.

- (b) Using `BRugs` and `WinBUGS`, perform a Bayesian logistic regression analysis of the intensive care unit (ICU) data from Question 1 of Assignment 6. Only use the predictors for the final model from that analysis: `age`, `cancer`, `SBP`, `emergency`, `hiPH`, `hiPCO2` and `coma`. You should use standardised versions of the continuous predictors `age` and `SBP`.

Hint: Modify code in the `bpdBayes*` files used in the BUGS preliminaries. Note Warnings I and II.