STAT40380/40390 - Bayesian Analysis

Computer Lab 3 – Simple statistical modelling and model comparison

1 WinBUGS data formats

So far we have only used the standard format for inputting data into WinBUGS. This involves creating a `list(...)` object which we load in as part of the model setup stage. This method can be quite awkward when data are presented in matrix format, or when there is a very long list of observations. Consider the model:

$$x_{ij} \sim N(\mu, \sigma^2),\ p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

where the data are given to us in the format:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{bmatrix}$$

(Note that the matrix format here is not strictly necessary, but we are using it to illustrate how to input this kind of data.)

There are two possibilities for entering this data into WinBUGS. One way is to use a `structure` command and include this in the list. An alternative is to use the vector data format also allowed by WinBUGS. The two versions are implemented in the files `matrix1.odc` and `matrix2.odc` on Blackboard. Note that this alternative format (found in `matrix2.odc`) requires you to click 'load data' twice; once after highlighting the list, and once again after highlighting the column names of the matrix-format data.

A word of warning: it is very easy to get matrices the wrong way round in WinBUGS as it reads them in oppositely to the computer software package R. Always check that your matrices are correctly imported into WinBUGS.

Task:

- Open the files `matrix1.odc` and `matrix2.odc`. Run each of the models and check that the posterior distributions of μ and σ are identical. The data were simulated with $\mu = 2$ and $\sigma = 1.5$. Do the results you have obtained seem consistent with these values?

2 Linear regression models

Where we have a response variable y and covariates $x_1, \ldots, x_p$ we may wish to fit a linear regression model, such that:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$. The model above defines a simple conditionally independent likelihood, and we may assume Jeffreys priors so that $p(\alpha, \beta_1, \ldots, \beta_p, \sigma^2) \propto \frac{1}{\sigma^2}$.

There are many extensions to such models. For example, polynomial regressions occur when $x_j = x^j$ for $j = 1, \ldots, p$. The parameters themselves might vary for each observation, creating a *hierarchical model* (covered later in the course), or the residual ϵ_i terms might not be normally distributed, which in some cases leads to a *generalised linear model.*

Tasks:

- Open and run the file `lr1.odc`. What model is being fitted here? Use the Inferences > Compare dialog box to produce a plot of the fitted values and data points. (Hint: watch the parameters `mu`, then run the model, then put `mu` in the 'node' box, `Y` in the 'other' box, and `X` in the 'axis' box.)
- Open and run the file `lr2.odc`. What model is being fitted here? Use the Inferences > Correlations dialog box to compare the correlations between the parameters `alpha` and `beta1`, and `alpha` and `beta2`. What is the difference between the two models? Why might one be preferred over the other?

3 Generalised linear models

In cases where the residuals ϵ are not normally distributed, but are a member of the exponential family, we can still model the mean (or more usually a function of the mean) as a linear function of explanatory variables. There are numerous examples:

- Poisson likelihood: $y_i \sim Po(\lambda_i)$, $\log(\lambda_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$
- Binomial likelihood: $y_i \sim Bin(n, p_i)$, $\text{logit}(p_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$
- Exponential likelihood: $y_i \sim Exp(\gamma_i)$, $\gamma_i^{-1} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$

In each case we must specify prior distributions on the parameters $\alpha, \beta_1, \ldots, \beta_p$. The transformation of the mean (eg log, logit, inverse, above) is known as the *link* function. The list above shows the most popular link functions for the specified models, though there are other choices that may be used.

Tasks:

- Open and run the file `glm11.odc`. The data are counts of mining disasters between 1851 and 1962. Which of the above models is being fitted here?
- Re-run the model, making sure to include `lambda` in the sampling box. Does this affect the running of your model? What use might the values of `lambda` be?
- What does the commented out line create? Remove the comment sign and re-run the model, monitoring the new variable `ystar`. What is the 95% credibility interval for `ystar`?

4 Comparing models: DIC

Later in the course we will meet a model comparison tool known as the *Deviance Information Criterion* or DIC. The DIC is similar to the AIC and BIC you may have met before. These tools allow for comparison between different models with differing structures and numbers of parameters, provided all models use the same data. The DIC is specifically built for Bayesian models and is calculated via the formula:

$$DIC = \bar{D} + p_D$$

where $\bar{D}$ is the mean deviance (a measure of model fit), and p_D is the effective number of parameters. We will discuss the DIC in more detail in lectures, but for now we will assume that lower DICs will lead to better models. The DIC is implemented in WinBUGS through the Inferences menu. Like other parameters, we need to 'set' it (and run the updates) in order to create an estimate of DIC for the model we are fitting.

Tasks:

- Open and run the files `glm1.odc` and `glm2.odc`. What is the difference between the models?
- Run each model and calculate the DIC. Which model is preferred? What are the values of p_D? Do they relate to the number of parameters in the model?

5 Homework: Putting distances

Some data are available on the number of successful putts from various distances for professional golfers. The data are as follows:

distance(feet)	no of tries	no of successes
x	n	y
2	1443	1346
3	694	577
4	455	337
5	353	208
6	272	149
7	256	136
8	240	111
9	217	69
10	200	67
11	237	75
12	202	52
13	192	46
14	174	54
15	167	28
16	201	27
17	195	31
18	191	33
19	147	20
20	152	24

Choose a suitable linear generalised linear model and fit the data in WinBUGS. Use DIC to choose between different models you might like to fit. Estimate the proportion of successes from 5, 10, and 30 feet. Write a short report of 2-3 pages with your final model code as an appendix.