MAS8303 Modern Bayesian Inference
Part 2

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

Chapter 2

Generalised Linear Models

2.1 Generalised Linear Models

2.1.1 Introduction

In this chapter of the course we are going to look at models which are more general than the normal linear model. There is not generally a conjugate form for the prior distribution so, except in simple cases where there are few parameters, we usually use Markov chain Monte Carlo (MCMC) methods to evaluate posterior distributions. In this course we shall use a R package called `rjags` which is an implementation of the "JAGS" ("Just Another Gibbs Sampler") system.

Consider the normal linear model. The i^{th} observation Y_i has a systematic component μ_i and a random component ε_i :

$$Y_i = \mu_i + \varepsilon_i.$$

We assume

- that ε_i has a normal distribution,
- that ε_i has variance σ^2,
- that ε_i is independent of ε_j for $i \neq j$.

In generalised linear models we relax the first two of these assumptions to allow a much wider class of models.

2.1.2 Linear predictors and link functions

In the normal linear model

$$\mu_i = \sum_{j=1}^{p} x_{i,j}\beta_j$$

where $\beta_1, \ldots, \beta_p$ are parameters and $x_{i,j}$ is the value of covariate j for observation i. Now we introduce a quantity called the *linear predictor:*

$$\eta_i = \sum_{j=1}^{p} x_{i,j}\beta_j.$$

In the normal linear model $\mu_i = \eta_i$. In a generalised linear model $\eta_i = g(\mu_i)$ where g is a known function called the *link function.* The link function must be monotonic and differentiable.

2.1.3 Error functions and the exponential family of distributions

In a generalised linear model the distribution of Y_i need not be normal. The mean is $\mathrm{E}(Y_i) = \mu_i$, where $\eta_i = g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j$, but the distribution may be chosen from a family of distributions, called the *exponential family,* which includes normal, binomial, Poisson and gamma. In some cases the variance of Y_i will depend on μ_i. E.g.

Normal	$N(\mu, \sigma^2)$	$\mathrm{var}(Y_i) = \sigma^2$	
Binomial	$\mathrm{Bin}(n, p)$	$\mathrm{var}(Y_i) = \mu(1 - \mu/n)$	$(\mu = np)$
Poisson	$\mathrm{Po}(\mu)$	$\mathrm{var}(Y_i) = \mu$	

In fact we could define models where the error distribution did not come from the exponential family but certain properties can be derived from the fact that the distribution does belong to the exponential family and so this is usually required for a model to qualify as a generalised linear model.

If a continuous random variable has an exponential family distribution then its density function has the form

$$f(y \mid \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\}.$$

If the variable is discrete rather than continuous then its probability function takes this form. The parameter θ is called the *canonical parameter.* The parameter ϕ is called the *scale parameter* and $\phi \geq 0$.

Normal distribution : $Y \sim N(\mu,\ \sigma^2)$.

$$\begin{aligned} f_Y(y) &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{y^2 - 2\mu y + \mu^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\} \\ &= \exp\left\{\frac{\mu y - \mu^2/2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Hence $\theta = \mu,\ \phi = \sigma^2,\ b(\theta) = \mu^2/2,\ a(\phi) = \phi = \sigma^2,\ c(y, \phi) = -(1/2)[y^2/\sigma^2 + \log(2\pi\sigma^2)]$.

Binomial distribution : $Y \sim \text{Bin}(n,\ p)$.

$$\begin{aligned}
f_Y(y) &= \binom{n}{y} p^y (1-p)^{n-y} \\
&= \binom{n}{y} \left(\frac{p}{1-p}\right)^y (1-p)^n \\
&= \exp\left\{\log\binom{n}{y} + y\log\left(\frac{p}{1-p}\right) + n\log(1-p)\right\} \\
&= \exp\left\{\log\binom{n}{y} + y\theta - n\log(1+e^\theta)\right\} \\
&= \exp\left\{\frac{y\theta - n\log(1+e^\theta)}{1} + \log\binom{n}{y}\right\}
\end{aligned}$$

So

$$\begin{aligned}
\theta &= \log\left(\frac{p}{1-p}\right) \\
e^\theta &= \frac{p}{1-p} \\
1+e^\theta &= 1 + \frac{p}{1-p} = \frac{1}{1-p} \\
\log(1+e^\theta) &= -\log(1-p) \\
b(\theta) &= n\log(1+e^\theta) \\
\phi = 1,\ a(\phi) &= 1 \\
c(y,\phi) &= \log\left(\binom{n}{y}\right)
\end{aligned}$$

2.1.4 Example

Consider the emission of α-particles by a radioactive source. We suppose that the emission rate at time t is $\beta e^{-\gamma t}$. Count the α-particles emitted in a short period of time, of length δt (short enough for the emission rate to be approximately constant) at each of $t_1, t_2, \ldots, t_n$ (equal periods at each). The mean number in a period of length δt at time t is $\delta t \beta e^{-\gamma t}$. Write this as $\exp(\beta_0 + \beta_1 t)$, where $\beta_0 = \ln(\delta t \beta)$ and $\beta_1 = -\gamma$. Suppose the actual number Y_i observed at time t_i has a Poisson distribution with mean $\mu_i = \exp(\beta_0 + \beta_1 t_i)$.

Hence the link function is log. The linear predictor is $\eta_i = \ln(\mu_i) = \beta_0 + \beta_1 t_i$. The error distribution is Poisson. So we have a generalised linear model.

2.1.5 Poisson Regression

Example 1

This example is based on a student project from some years ago. The project was conducted in collaboration with the Sunderland and South Shields Water Company. (It was *many* years ago!).

A water company has many kilometres of water pipe. Much of this lies under roads etc. Some of the pipes may be very old. From time to time bursts, fractures and leaks of various sorts occur. The company wants to investigate how the rate of failures depends on various factors and covariates. These might include age, diameter, material, depth below surface, number of customers supplied, whether the pipe is in a residential or industrial area etc. Some sections of pipe have been observed for longer than others.

We assume that, for a given section of pipe, the rate at which failures occur is proportional to the length of the section. We also assume that, over relatively short periods compared to the lifetime of a pipe, e.g. a year, the rate remains more or less constant and the actual number of failures has a Poisson distribution with mean proportional to the length of the period. So, for a particular section of pipe, of length k_i, observed over a period of length t_i, the mean number of failures would be $\mu_i = \lambda_i k_i t_i$. The parameter λ_i depends on the covariates for that pipe (age, diameter etc.) in the period in question. The mean of a Poisson distribution has to be positive. This can be ensured if we use a log link function so that $\ln(\mu_i) = \ln(\lambda_i) + \ln(k_i) + \ln(t_i)$. Now we apply a linear model for $\eta_i = \ln(\lambda_i)$. We could include the terms in $\ln(k_i)$ and $\ln(t_i)$ in the linear model but we know the values of the coefficients of these (i.e. 1). The other covariates have unknown coefficients and we need to give these a prior distribution. Thus

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j}$$

or, in matrix notation,

$$\underline{\eta} = X\underline{\beta}.$$

This is thus a generalised linear model with Poisson errors and log link. We might well give a multivariate normal prior distribution to the unknown β coefficients.

Example 2

Patients in four groups are observed for various lengths of time. During this time tumours may develop. The dependent variable is the number of tumours observed for each patient. The mean number of tumours for a patient in group g is $\lambda_g t = \exp(\beta_g + \ln t)$ where t is the time observed in weeks. Thus the parameters are $\beta_1, \ldots, \beta_4$ where $\beta_g = \ln(\lambda_g)$. There is no intercept here and the coefficient of $\ln t$ is known to be 1. If we included an intercept then we would have to drop one of the group parameters, exactly as in linear models.

Using BRugs

We will be using `rjags` to do practical work. This is a R package which implements a Gibbs sampler. Models and priors are specified using the a model specification language which is essentially the BUGS ("Bayesian inference Using Gibbs Sampling") language. We will need to make a *model*

```
model

{
 for (i in 1:N)
     {y[i]~dpois(mean[i])
      mean[i]<-lambda[group[i]]*t[i]
      }

 for (g in 1:4)
      {lambda[g]<-exp(beta[g])
       beta[g]~dnorm(mu,10)
       }
 mu~dnorm(0,5)
 }
```

Figure 2.1: BUGS code for the tumours example

specification file using the BUGS language. As an example, consider Example 2 above. Suppose that we observe n patients (written as `N` in the BUGS code). For each patient we have a group number g (`group`), the time t for which the patient was observed (`t`) and the number y of tumours observed (`y`).

Figure 2.1 shows some suitable BUGS code. Note that this is *not* a program with commands to be executed. It is a model specification. We are defining the joint distribution of the unknowns and the data, mostly by specifying conditional distributions. For example

```
y[i]~dpois(mean[i])
```

might be written in standard mathematical notation as

$$Y_i \mid m_i \sim \text{Po}(m_i).$$

The code `dpois` represents the Poisson distribution and the symbol `~` has its usual meaning of "has the following distribution." Similarly `dnorm` stands for a normal distribution. Note however that the parameters are mean and *precision*, not mean and variance. So we are saying that

$$\beta_g \mid \mu \sim N(\mu,\ 0.1).$$

Notice that we are giving $\beta_1, \ldots, \beta_4$ a *hierarchical* normal prior. Since

$$\mu \sim N(0,\ 0.2),$$

the prior mean of β_g is 0, the prior variance of β_g is $0.2 + 0.1 = 0.3$ but $\beta_1, \ldots, \beta_4$ are not independent in the prior. We have $\text{covar}(\beta_g, \beta_{g'}) = \text{var}(\mu) = 0.2$ when $g \neq g'$.

2.2 Binomial Regression

2.2.1 Introduction

Just as we can have a regression where the error distribution is Poisson we can have a regression where the error distribution is binomial.

The term "logistic regression" is often used. Strictly this should refer to cases where the logistic link function is used. There are other suitable link functions.

Suppose, for example, we want to know what factors influence whether or not a person will buy a particular product. We might have data on a number of variables, such as age, sex, marital status, income, etc. and, of course, whether or not they buy the product, for each of a sample of individuals. The response variable here is binary. That is $y_i = 1$ if person i buys the product and otherwise $y_i = 0$. We can think of the mean of y_i as p_i, the probability that an individual with the same covariate values as individual i would buy the product. A regression model would relate p_i to the values of the explanatory variables. Clearly a linear model $p_i = \sum \beta_j x_{ij}$ is inappropriate since large values of $\sum \beta_j x_{ij}$ would lead to fitted values of p_i greater than 1 and small values of $\sum \beta_j x_{ij}$ would lead to fitted values of p_i less than 0. Instead we transform p_i from a $(0, 1)$ scale to a $(-\infty, \infty)$ scale. This is usually done using a *sigmoid*, i.e. S-shaped function. The transformation which gives logistic regression its name is the logistic transformation. The transformed proportions are sometimes called *logits.*

$$\eta_i = \ln\left\{\frac{p_i}{1-p_i}\right\}.$$

Notice that if $p_i \to 1$ then $\eta_i \to \infty$ and if $p_i \to 0$ then $\eta_i \to -\infty$.
The inverse transformation is

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

Another popular transformation is "probits"

$$\begin{aligned} \eta_i &= \Phi^{-1}(p_i), \\ p_i &= \Phi(\eta_i), \end{aligned}$$

where $\Phi()$ is the standard normal distribution function and $\Phi^{-1}()$ is its inverse.
Yet another is the complementary log-log link,

$$\begin{aligned} \eta &= \ln[-\ln(1-p)], \\ p &= 1-\exp(-e^{\eta}). \end{aligned}$$

```
model

{
 for (i in 1:7)
     {effects[i]~dbin(p[i],n[i])
      logit(p[i])<-beta2+beta*(dose[i]-2)
      }

 alpha<-beta2-2*beta

 beta2~dnorm(-0.27, 2.17)
 beta~dnorm(0.81, 8.61)
 }
```

Figure 2.2: BUGS code for the side-effect example.

2.2.2 Example

The proportion of people, given a drug to treat a medical condition, who contract a particular side effect depends on the dose of the drug. If p is the proportion suffering the side effect at dose x, then

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

where α and β are parameters with unknown values.

At each of a number of doses, x_i, a number, n_i, of patients were given the drug and the number, r_i, with the side effect was recorded.

Dose	x_i	0.9	1.1	1.8	2.3	3.0	3.3	4.0
No. patients	n_i	46	72	118	96	84	53	38
No. with side effect	r_i	17	22	52	58	56	43	30

2.2.3 Using JAGS

Figure 2.2 shows some BUGS code for the example above.

Notice that `effects[i]~dbin(p[i],n[i])` says that

$$r_i \mid p_i \sim \text{binomial}(n_i, p_i).$$

That is, the BUGS notation for binomial distributions is the other way round to the usual convention in this case. Notice also that we are allowed to put the function `logit(p[i])` on the left of `<-` . On the right of this statement we have $\beta_2 + \beta(x_i - 2)$. This is to illustrate what we can do in a regression when the intercept is not a convenient quantity for prior specification. Here it is supposed to be more convenient to think about the rate of side effects when the dose is 2 rather than when it is zero. See below.

2.2.4 Prior specification

In the example above we suppose that we are prepared to consider the probability of a side effect when the dose is 2. Denote this probability π_2. The Bayesian statistics literature includes the results of careful research into how best to *elicit* a prior distribution for a probability such as this. Unfortunately we do not have time to go into detail. Suppose that, in our prior beliefs, we assess $\Pr(\pi_2 < 0.2) = \Pr(\pi_2 > 0.7) = 0.05$. Let $\beta_2 = \log(\pi_2/(1-\pi_2))$. Then we believe that

$$\begin{aligned} \Pr\left[\beta_2 < \log\left(\frac{0.2}{1-0.2}\right) = -1.3863\right] &= 0.05, \\ \Pr\left[\beta_2 > \log\left(\frac{0.7}{1-0.7}\right) = 0.8473\right] &= 0.05 \end{aligned}$$

Now suppose that we give β_2 a normal $N(\mu,\ \sigma^2)$ prior distribution. From the properties of the normal distribution we deduce that

$$\begin{aligned} \mu - 1.645\sigma &= -1.3863, \\ \mu + 1.645\sigma &= 0.8473. \end{aligned}$$

this leads to $\mu = (-1.3863 + 0.8473)/2 = -0.2695$ and $\sigma = (0.8473 - [-1.3863])/(2 \times 1.645) = 0.6789$ and therefore $\sigma^2 = 0.4609$.
This gives us a prior distribution for β_2. It is normal with mean -0.2695 and precision $1/0.4609 = 2.1696$. It does not seem unreasonable to round these to -0.27 and 2.17 in this case. So, we have a prior for one point on the regression line. We need a prior for the gradient. Suppose that we are willing to give $\gamma = \beta_4 - \beta_2$ a normal prior, independently of β_2, where $\beta_4 = \log[\pi_4/(1-\pi_4)]$ and π_4 is the probability of a side effect when the dose is 4. Suppose that, by a process similar to that for β_2 we assign a $N(0.6750,\ 0.8563)$ distribution to β_4. (Start with $\Pr(\pi_4 < 0.3) = \Pr(\pi_4 > 0.9) = 0.05$). Then, since $\beta_4 = \beta_2 + \gamma$ and β_2 and γ are independent, we can deduce that $\gamma \sim N(0.9445,\ 0.3954)$. However $\gamma = \beta(4-2) = 2\beta$ so our prior distribution for β becomes $N(0.4723,\ 0.0989)$ or, after rounding, normal with mean 0.47 and precision 10.12. (Many people might prefer a weaker prior distribution).

2.3 Log-linear Models for Categorical Data

2.3.1 Introduction

In this section we give a brief introduction to the analysis of categorical data using log-linear models. This is a large and complicated topic and we only scratch the surface here. An important special case is the analysis of contingency tables.

Suppose we have a single sample where each individual is classified into one of K categories. Associated with each individual is a vector of covariates and the probability of the individual being in each category depends on the covariates. For example, the categories might be the possible parties for which an individual will vote in an election. The covariates might be things like sex, age-group, occupation, usual newspaper. It may be that we can observe more than one individual with exactly the same covariates (e.g. women aged 20-29 who are students and read the Guardian). So, in this case, we can think of an "observation" as refering to a group of individuals who have the same covariate values. Let group i refer to the individuals with covariate pattern x_i. Suppose that there are I such groups. Let the number in group i be N_i (which might be 1, of course) and let the number of these who are observed to be in category k (e.g. vote for party k) be $n_{i,k}$. Let $\underline{n}_i = (n_{i,1}, \ldots, n_{i,K})^T$. The appropriate distribution for $\underline{n}_i$ is the multinomial distribution and the likelihood is as follows where the probability for category k given covariate pattern x_i is $p_{i,k}$,

$$\sum_{k=1}^{K} n_{i,k} = N_i \qquad \text{and} \qquad \sum_{k=1}^{K} p_{i,k} = 1.$$

The likelihood is

$$L = \prod_{i=1}^{I} \frac{N_i! p_{i,1}^{n_{i,1}} p_{i,2}^{n_{i,2}} \cdots p_{i,K}^{n_{i,K}}}{n_{i,1}! n_{i,2}! \cdots n_{i,K}!}.$$

Let $\mu_{i,k} = N_i p_{i,k}$. Since $\sum_k p_{i,k} = 1$ we have $\sum_k \mu_{i,k} = N_i$. Now we can write the likelihood as follows.

$$\begin{aligned}
L &= \prod_{i=1}^{I} \frac{N_i! (\mu_{i,1}/N_i)^{n_{i,1}} (\mu_{i,2}/N_i)^{n_{i,2}} \cdots (\mu_{i,K}/N_i)^{n_{i,K}}}{n_{i,1}! n_{i,2}! \cdots n_{i,K}!} \\
&= \prod_{i=1}^{I} \frac{N_i!}{N_i^{N_i}} \prod_{k=1}^{K} \frac{\mu_{i,k}^{n_{i,k}}}{n_{i,k}!} \\
&= \prod_{i=1}^{I} \frac{N_i!}{N_i^{N_i}} \exp\left(\sum_{k=1}^{K} \mu_{i,k}\right) \prod_{k=1}^{K} \frac{e^{-\mu_{i,k}} \mu_{i,k}^{n_{i,k}}}{n_{i,k}!} \\
&= \prod_{i=1}^{I} \frac{N_i!}{N_i^{N_i}} e^{N_i} \prod_{k=1}^{K} \frac{e^{-\mu_{i,k}} \mu_{i,k}^{n_{i,k}}}{n_{i,k}!}
\end{aligned}$$

Thus the likelihood is proportional to that for Poisson data.

To complete the generalised linear model we need an appropriate link function. One way to do this is to set

$$p_{i,k} = \frac{e^{\eta_{i,k}}}{\sum_{k'} e^{\eta_{i,k'}}} \tag{2.1}$$

and

$$\eta_{i,k} = \sum_{j=1}^{J} \beta_{j,k} x_{i,j}$$

where $x_{i,j}$ is the value of covariate j in pattern i.

However, looking at (2.1) we see that the parameters are not *identifiable*. This is because we can write

$$\eta_{i,k} = \sum_{j=1}^{J} \beta_{j,k} x_{i,j} = \sum_{j=1}^{J} (\beta_{j,k} - \beta_{j,1}) x_{i,j} + \sum_{j=1}^{J} \beta_{j,1} x_{i,j}.$$

Now write $\tilde{\beta}_{j,k} = \beta_{j,k} - \beta_{j,1}$ and

$$\tilde{\eta}_{i,k} = \sum_{j=1}^{J} \tilde{\beta}_{j,k} x_{i,j} = \eta_{i,k} - \sum_{j=1}^{J} \beta_{j,1} x_{i,j}.$$

If we substitute $\tilde{\eta}_{i,k}$ for $\eta_{i,k}$ in (2.1) we get exactly the same value for $p_{i,k}$. Therefore, without loss of generality in terms of the likelihood we can set $\beta_{1,1} = \cdots = \beta_{J,1} = 0$ and therefore $\eta_{i,1} = 0$ and $\exp(\eta_{i,1}) = 1$. Then (2.1) is equivalent to

$$\ln\left(\frac{p_{i,k}}{p_{i,1}}\right) = \ln\left(\frac{\mu_{i,k}}{\mu_{i,1}}\right) = \sum \beta_{j,k} x_{i,j}$$

for $k = 2, \ldots, K$. We do not need to apply this model to $p_{i,1}$ since we know that $\sum p_{i,k} = 1$.

Of course we need not pick the first category as the baseline. We could pick any. Also, although this constraint makes no difference to the likelihood, it may make specification of the prior a little awkward. An alternative constraint is to set

$$\sum_{k=1}^{K} \beta_{j,k} = 0.$$

2.3.2 Example

The following data are taken from Freeman (1987). Babies were categorised as follows.

1 Full term, alive at end of year 1.

2 Full term, died in first year.

3 Premature, alive at end of year 1.

4 Premature, died in first year.

The mothers were categorised as either "Young" or "Older" as as either "Smokers" or "Non-smokers." Interest lies in the effects of the mother's age and smoking on the outcome.

Mother		Outcome				
Age	Smoking	1	2	3	4	Total
Young	Non-smoker	4012	24	315	50	4401
Young	Smoker	459	6	40	9	514
Older	Non-smoker	1594	14	147	41	1796
Older	Smoker	124	1	11	4	140

In this case it is natural to use Category 1 as a baseline since this is the "normal" outcome and we are interested in the risks of the other outcomes. For the other three categories, $k = 2, 3, 4$, we can model $\eta_{i,k}$ as follows.

Young, Non-smoker	$\eta_{1,k} = \beta_{0,k} - \beta_{a,k} - \beta_{s,k} + \beta_{as,k}$
Young, Smoker	$\eta_{2,k} = \beta_{0,k} - \beta_{a,k} + \beta_{s,k} - \beta_{as,k}$
Older, Non-smoker	$\eta_{3,k} = \beta_{0,k} + \beta_{a,k} - \beta_{s,k} - \beta_{as,k}$
Older, Smoker	$\eta_{4,k} = \beta_{0,k} + \beta_{a,k} + \beta_{s,k} + \beta_{as,k}$

Here $\beta_{a,k}$ is an age effect, $\beta_{s,k}$ is a smoking effect and $\beta_{as,k}$ is an interaction effect between age and smoking. In effect we have a covariate "age" which takes the values $(-1, -1, 1, 1)$ in the four groups and so on.

Now we need a prior distribution for these β coefficients. We could spend more time looking at this in detail but here is something fairly simple.

$$\begin{array}{rr} \beta_{0,k} \mid \mu_0 \sim N(\mu_0,\ 1.0) & \mu_0 \sim N(-2,\ 1.0) \\ \beta_{a,k} \mid \mu_a \sim N(\mu_a,\ 0.1) & \mu_a \sim N(0,\ 0.1) \\ \beta_{s,k} \mid \mu_s \sim N(\mu_s,\ 0.1) & \mu_s \sim N(0,\ 0.1) \\ \beta_{as,k} \mid \mu_{as} \sim N(\mu_{as},\ 0.05) & \mu_{as} \sim N(0,\ 0.05) \end{array}$$

In each case we have used a "hierarchical" prior so that, e.g., $\beta_{0,2}, \beta_{0,3}, \beta_{0,4}$ are correlated in the prior.

Figure 2.3 shows some suitable BUGS code.

2.3.3 Contingency tables

Suppose we have a (2-dimensional) contingency table with R rows and C columns. This could arise in two quite different ways:

1. It could be the result of taking a single sample of individuals and categorising them in two ways (e.g. by occupation and by which newspaper they read).

2. Each row might be a separate sample and the individuals are categorised according to the column classification (e.g. we take a sample from each of several occupations and ask which newspaper each person reads).

Although, in non-Bayesian statistics, the same χ^2 test is applied in both cases, the two situations are really quite different and the Bayesian analyses of them are different. In this section we will be looking at case 1 only. This is really a special case of the loglinear models already discussed where there are no covariates but we parameterise the multinomial distribution in terms of the row and column factors. So the probability of an observation falling into the row r, column c cell may depend on a row effect, a column effect and, possibly, a row-column interaction effect. If we include both the main effects and the interaction effect then we have a *saturated* model with the maximum number of parameters. We may be interested in looking at the posterior distribution of the interaction effect to see whether there is evidence of dependence between the row and column categorisations.

2.3.4 Example

The following data are taken from Krzanowski (1988). Schoolchildren were examined and classified according to the size of their tonsils and whether or not they were carriers of the bacterium *Streptococcus pyogenes*. In total 1398 children were examined.

```
model

{
 for (i in 1:4)
     {y[i,1:4]~dmulti(p[i,],n[i])
      for (k in 1:4)
          {p[i,k]<-phi[i,k]/sum(phi[i,])
           phi[i,k]<-exp(eta[i,k])
           }

      for (k in 1:4)
          {eta[i,k]<-beta0[k]+betaa[k]*age[i]+betas[k]*smoke[i]+betaas[k]*age[i]*smoke[i]
           }
      }

 beta0[1]<-0
 betaa[1]<-0
 betas[1]<-0
 betaas[1]<-0

 for (k in 2:4)
     {beta0[k]~dnorm(mu0,1.0)
      betaa[k]~dnorm(mua,10.0)
      betas[k]~dnorm(mus,10.0)
      betaas[k]~dnorm(muas,20.0)
      }

 mu0~dnorm(-2,1.0)
 mua~dnorm(0,10.0)
 mus~dnorm(0,10.0)
 muas~dnorm(0,20.0)

 }
```

Figure 2.3: BUGS code for Example 2.3.2

	Carrier status	
Tonsil size	Carrier	Non-carrier
Normal	19	497
Large	29	560
Very large	24	269

Of course we could just give the six probabilities a Dirichlet prior but another possibility is to parameterise the model as follows.

Carrier	Normal	$\eta_{1,1} =$	β_1	$-2\beta_2$		$-2\beta_4$	
Carrier	Large	$\eta_{2,1} =$	β_1	$+\beta_2$	$-\beta_3$	$+\beta_4$	$-\beta_5$
Carrier	Very large	$\eta_{3,1} =$	β_1	$+\beta_2$	$+\beta_3$	$+\beta_4$	$+\beta_5$
Non-carrier	Normal	$\eta_{1,2} =$	$-\beta_1$	$-2\beta_2$		$+2\beta_4$	
Non-carrier	Large	$\eta_{2,2} =$	$-\beta_1$	$+\beta_2$	$-\beta_3$	$-\beta_4$	$+\beta_5$
Non-carrier	Very large	$\eta_{3,2} =$	$-\beta_1$	$+\beta_2$	$+\beta_3$	$-\beta_4$	$-\beta_5$

Notice that, whatever the values of $\beta_1, \ldots, \beta_5$, if we sum $\eta_{1,1}, \ldots, \eta_{3,2}$, we always get zero. Notice also that

- β_1 is a carrier effect
- β_2 is a large-tonsil effect
- β_3 is a very-large-tonsil effect
- β_4 and β_5 are interaction effects. The coefficients of β_4 are obtained by multiplying those of β_1 and β_2. The coefficients of β_5 are obtained by multiplying those of β_1 and β_3.

The particular structure which we have here reflects the fact that "Normal", "Large", "Very large" are *ordered categories*.

Slightly adapting (2.1), we now set

$$p_{i,j} = \frac{e^{\eta i,j}}{\sum\sum e^{\eta_{i,j}}}.$$

To find a suitable prior distribution for each of the β parameters we need to think about log odds, for example the log of the probability of being a carrier divided by the probability of being a non-carrier. We will omit the details and use the following independent priors.

$$\begin{array}{cc} \beta_1 \sim N(-1.5, 2.5) & \beta_2 \sim N(0, 1.6) \\ \beta_3 \sim N(0, 1.6) & \beta_4 \sim N(0, 1.0) \\ \beta_5 \sim N(0, 1.0) & \end{array}$$

Figure 2.4 shows some suitable BUGS code.

Notice that we have arranged the ηs into a single vector for convenience. Notice also that some extra quantities are calculated at the end. This is simply so that we can easily find the posterior distributions of these quantities. Let R_{normal} be the conditional probability of being a carrier given normal-sized tonsils, and similarly R_{large} and R_{vlarge} for large and very large tonsils. Then we calculate two log relative risks: $\log(R_{\text{large}}/R_{\text{normal}})$ and $\log(R_{\text{vlarge}}/R_{\text{normal}})$ to see how much enlarged tonsils affects the probability of a child being a carrier.

```
model

{
 y[1:6]~dmulti(p[],1398)

 for (k in 1:6)
     {p[k]<-phi[k]/sum(phi[])
      phi[k]<-exp(eta[k])
           }

  eta[1]<-  beta[1]-2*beta[2]        -2*beta[4]
  eta[2]<-  beta[1]+  beta[2]-beta[3]+  beta[4]-beta[5]
  eta[3]<-  beta[1]+  beta[2]+beta[3]+  beta[4]+beta[5]
  eta[4]<- -beta[1]-2*beta[2]        +2*beta[4]
  eta[5]<- -beta[1]+  beta[2]-beta[3]-  beta[4]+beta[5]
  eta[6]<- -beta[1]+  beta[2]+beta[3]-  beta[4]-beta[5]

  beta[1]~dnorm(-1.5,0.4)
  beta[2]~dnorm(0,0.625)
  beta[3]~dnorm(0,0.625)
  beta[4]~dnorm(0,1.0)
  beta[5]~dnorm(0,1.0)

  rnormal<-p[1]/(p[1]+p[4])
  rlarge<-p[2]/(p[2]+p[5])
  rvlarge<-p[3]/(p[3]+p[6])

  lrrlarge<-log(rlarge/rnormal)
  lrrvlarge<-log(rvlarge/rnormal)

 }
```

Figure 2.4: BUGS code for tonsils example

2.4 Practical 2

2.4.1 Introduction

In this practical we will start to use the R package `rjags` to do MCMC evaluation of posterior distributions. We will do some examples involving generalised linear models.

The software BUGS (**B**ayesian Inference **U**sing **G**ibbs **S**ampling) was developed to allow users to specify models and priors, connect these with data and compute samples of unknowns from the posterior distribution using a Gibbs sampler (Spiegelhalter *et al*, 1995). Later a menu-driven version to run under MS Windows, called WinBUGS (Lunn *et al*, 2000) was developed. This eventually incorporated new features not found in the original, or 'Classic", BUGS. There are now also OpenBUGS, developed at the University of Helsinki, JAGS (**J**ust **A**nother **G**ibbs **S**ampler) (Plummer, 2012) and various other implementations of the basic "BUGS" idea. In particular we will be using `rjags` which is a R package which implements JAGS within R. All of these use (apart from a few small differences) the same *Model Specification Language* and, in this part of the module, it is this language, and model specification generally, which are of particular interest.

The WinBUGS manual is available from the MAS8303 Web Page. The details of how you tell `rjags` to do things are different from WinBUGS but the model specification language and many other features are the same.The JAGS and `rjags` manuals are available from Dr Farrow's MAS8391 Web page at

`http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8391/`

Henceforth we will refer to this Web page as "MF's Web Page".

2.4.2 Loading `rjags`

Start R. You may well wish to change the working directory, for example to a MAS8303 folder. This can be done via the *File Menu*.

Type:

```
library(rjags)
```

2.4.3 Poisson regression: Aircraft fatalities

This example has only two parameters so we do not really *need* MCMC but it will serve as a first example.

The data in table 2.1 come from Phillips (1978). People sometimes commit "murder-suicide" by deliberately crashing private aircraft. It was thought that newspaper coverage of such an event might trigger other incidents. The data give the number of "multi-fatality crashes" in the week following each of 17 known cases of murder-suicide, together with an index of newspaper coverage. The idea is to investigate whether the number of crashes is related to the newspaper coverage.

We adopt the following model.

$$\begin{aligned} Y_i \mid \beta_0, \beta_1 &\sim \text{Po}(\mu_i) \\ \eta_i = \log(\mu_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

We give the two parameters independent priors as follows.

$$\begin{aligned} \beta_0 &\sim N(1, 4) \\ \beta_1 &\sim N(0, 0.0001) \end{aligned}$$

1. Obtain the data file from MF's Web Page. Save the file as `aircraftdata.txt`.

2. Create a file called `aircraftbug.txt` containing the model specification as follows. You can use *Notepad* to do this.

```
model

{
 for (i in 1:17)
     {y[i]~dpois(mu[i])
      log(mu[i])<-beta0+beta1*x[i]
      }

 beta0~dnorm(1,0.25)
 beta1~dnorm(0,10000)
 }
```

3. Read the data into R and put them in a suitable format.

```
aircraft<-read.table("aircraftdata.txt",header=TRUE)
aircraftdata<-list(x=aircraft$x,y=aircraft$y)
```

4. Create a JAGS model object.

```
aircraftjags<-jags.model("aircraftbug.txt",data=aircraftdata,n.chains=2)
```

   Note that there is an argument which is the number of parallel chains which we want to use. Using parallel chains can be useful for checking convergence. Here we are using two chains. We can also specify initial values if we so wish.

5. Run the sampler for a burn-in period (of 5000 iterations here).

```
update(aircraftjags,5000)
```

6. Run the sampler for 10000 more iterations, recording the samples.

```
aircraftsamples<-coda.samples(aircraftjags,c('beta0','beta1'),10000)
```

7. At this stage we can check convergence of the chain by looking at a *trace plot*. Before we ask for the plots, it is advisable to change one of the R graphics parameters. We then have to click on the graphics window to move to the next plot.

```
par(ask=TRUE)
traceplot(aircraftsamples)
```

8. If we are satisfied that the chains had reached convergence (close enough) when we started to record samples, we can now look at some summaries of the posterior distribution.

```
summary(aircraftsamples)
```

9. We can also find approximations to the marginal posterior densities of the parameters.

```
densplot(aircraftsamples)
```

   We might want to do more sophisticated things such as change the way the density estimate is calculated or make a contour plot of the joint posterior distribution of the two parameters. To do these things we can extract the MCMC samples themselves and then do whatever we like with them. For example

```
aircraftsamplesout<-as.matrix(aircraftsamples,iters=TRUE)
```

   puts all of the recorded sampled values of β_0 and β_1 into the matrix `aircraftsamplesout`, along with the iteration numbers.

x	y	x	y	x	y
376	8	96	8	5	3
347	5	85	6	5	2
322	8	82	4	0	4
104	4	63	2	0	3
103	6	44	7	0	2
98	4	40	4		

Table 2.1: Index of newspaper coverage x and number of multi-fatality crashes y in weeks following incidenst of murder-suicide.

2.4.4 Binomial regression

This is the example in section 2.2.2. Use a similar procedure to that for the Poisson regression above. You will need to put the model specification into a file. You can also put the data into a file which should look like this.

```
 dose       n    effects
  0.9      46         17
  1.1      72         22
  1.8     118         52
  2.3      96         58
  3.0      84         56
  3.3      53         43
  4.0      38         30
```

Alternatively you can simply define the variables directly in R, *eg*

```
dose<-c(0.9,1.1,1.8,2.3,3.0,3.3,4.0)
```

and then, *eg*

```
sideeffect<-list(dose=dose,n=n,effects=effects)
```

2.4.5 Loglinear models: Babies

This is the example in section 2.3.2. Use a similar procedure to that for the Poisson regression above. The model specification is available from MF's Web page. It is a good idea to put the data into a file. The data file might look like this.

```
  y1       y2       y3      y4       n   age  smoke
4012       24      315      50    4401    -1      1
 459        6       40       9     514    -1     -1
1594       14      147      41    1796     1     -1
 124        1       11       4     140     1      1
```

You could then use something like

```
babies<-read.table("babies.txt",header=TRUE)
y<-with(babies,cbind(y1,y2,y3,y4))
babies<-list(y=y,n=babies$n,age=babies$age,smoke=babies$smoke)
```

Which quantities do you think that you should monitor (ie. record samples)? We could use, for example,

```
babysamples<-coda.samples(babyjags,c("beta0","betaa","betas","betaas"),10000)
```

and then, later,

```
traceplot(babysamples)
```

etc.Note that, in order for this `coda.samples` command to work, it was necessary to define `beta0[1]`, `betaa[1]`, `betas[1]` and `betaas[1]` in the model specification even though we do not really need them.

2.4.6 Loglinear models: Tonsils

This is the example in section 2.3.4. Use a similar procedure to that for the Poisson regression above. The model specification is available from MF's Web page. You can easily specify the data directly in R as follows.

```
tonsilsdata<-list(y=c(19,29,24,497,560,269))
```

Which quantities do you think that you should monitor?

2.5 Exercises

1. Observations are made on the numbers of caterpillars on commercially grown cabbages in J plots. The number of observations in plot j is n_{ij}. Let the number of caterpillars on the i^{th} cabbage in plot j be Y_{ij}. Given the values of $\lambda_1, \ldots, \lambda_J$, we have

$$Y_{ij} \mid \lambda_j \sim \text{Po}(\lambda_j),$$

a Poisson distribution with mean λ_j, and $Y_{11}, \ldots, Y_{n_J J}$ are conditionally independent.

Let $\eta_j = \log(\lambda_j)$. Given the values of μ and τ, we have

$$\eta_j \mid \mu, \tau \sim N(\mu,\ \tau^{-1}),$$

a normal distribution with mean μ and precision τ, and $\eta_1, \ldots, \eta_J$ are conditionally independent.

We have independent prior distributions for μ and τ with $\mu \sim N(m, v)$ and $\tau \sim \text{gamma}(a, b)$.

We make observations $Y_{ij} = y_{ij}$ and wish to use a Gibbs sampler to evaluate the posterior distribution.

Find a function proportional to the density of the full conditional distribution of η_j.

2. A particular surgical operation performed on patients with a serious condition is hazardous and a proportion of the patients die during surgery. Researchers wish to investigate the relationship between the death rate and the age of the patient. We have the following model. Let θ_x be the death rate for patients aged x years. That is, given θ_x, the probability of death is θ_x. Let

$$\eta_x = \log\left(\frac{\theta_x}{1-\theta_x}\right).$$

We suppose that

$$\eta_x = a + bx$$

for some unknown parameters $a,\ b$.

We develop a prior distribution as follows. Consider two ages, $x = 50$ and $x = 70$. Our marginal prior distributions for η_{50} and η_{70} are $\eta_{50} \sim N(m_{50}, v_{50})$ and $\eta_{70} \sim N(m_{70}, v_{70})$. The prior correlation of η_{50} and η_{70} is 0.8. We assess

$$\Pr(\theta_{50} < 0.05) = \Pr(\theta_{50} > 0.20) = \Pr(\theta_{70} < 0.1) = \Pr(\theta_{70} > 0.4) = 0.025.$$

(a) Find the values of $m_{50},\ v_{50},\ m_{70},\ v_{70}$ and the covariance of η_{50}, η_{70}.

(b) Find the joint prior distribution of a, b.

3. In an experiment on student learning, randomly selected students are assigned to groups which are given different amounts of tuition. Suppose group i has n_i students who are given $t_i + 30$ hours of tuition.

   At the end of the experiment the students are given a test. Suppose that a student's percentage mark is Z. Let $X = \ln(Z)$. Suppose that, for a student in group i, we assume $X \sim N(\alpha + \beta t_i, \sigma^2)$, where $\sigma = 0.1$. Instead of the actual percentage marks, all that is recorded is whether each student passes or fails the test. A student passes if $Z \geq 40$, that is $X \geq \ln 40$.

   Let y_i be the number of students in group i who pass the test.

   (a) Express this model as a generalised linear model.

   (b) State the link function and error function.

   (c) Find the linear predictor.

   (d) Use BRugs to evaluate the posterior distributions of α and β. You may use independent priors for α and β with

   $$\alpha^* \sim N(0.1, 0.01), \quad \alpha = \alpha^* + \ln 40 \quad \text{and} \quad \beta \sim N(0.0, 0.0004).$$

   The data are as follows.

t_i	n_i	y_i
-10	30	19
0	40	30
10	30	27

   (e) What happens if we do not assume that $\sigma = 0.1$ but allow σ^2 to be unknown?