MAS8303 Modern Bayesian Inference
Part 2

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

Chapter 0

Inference for More Than One Unknown

0.1 More than one unknown

0.1.1 Basic ideas

MAS3301 mostly looked at Bayesian inference in the case where we have a single unknown quantity, usually a parameter. In MAS8303 we will typically look at models with two or more, sometimes many more, unknowns. So, in this lecture, we will look at what happens when we have more than one unknown parameter. The principle is the same when we have more than one parameter. We simply obtain a joint posterior distribution for the parameters. For example, if there are two parameters, we might produce a contour plot of the posterior pdf, as shown in figure 1, or a "3-d" plot, as shown in figure 2. If there are more than two parameters we need to "integrate out" some of the parameters in order to produce graphs like this.

As usual, the basic rule is **posterior** $\propto$ **prior** $\times$ **likelihood.** If necessary, the normalising constant is found by integrating over all parameters. Posterior means, variances, marginal probability density functions, predictive distributions etc. can all be found by suitable integrations. In practice the integrations are often carried out numerically by computer. Apart from being the only practical means in many cases, this removes the pressure to use a convenient conjugate prior.

Sometimes our beliefs might be represented by a model containing several parameters and we might want to answer questions about a number of them. For example, in a medical experiment, we might be interested in the effect of a new treatment on several different outcome measures so we might want to make inferences about the change in the mean for each of these when we move from the old to the new treatment. In frequentist statistics this can give rise to the "multiple testing problem." This problem does not arise for Bayesians. For a Bayesian the inference always consists of the posterior distribution. Once we have calculated the posterior distribution we can calculate whatever summaries we want from it without any logical complications. For example, we could calculate a posterior probability that the mean outcome measure has increased from one treatment to the other for each outcome, or a joint probability that it has increased for every member of some subset of the outcomes or any or all of many other summaries.

0.1.2 The bivariate normal distribution

The normal distribution can be extended to deal with two variables. (In fact, we can extend this to more than two variables).

If Y_1 and Y_2 are two continuous random variables with joint pdf

$$f(\underline{y}) = (2\pi)^{-1}|V|^{-1/2}\exp\left\{-\frac{1}{2}(\underline{y}-\underline{\mu})^T V^{-1}(\underline{y}-\underline{\mu})\right\}$$

for $-\infty < y_1 < \infty$ and $-\infty < y_2 < \infty$ then we say that Y_1 and Y_2 have a bivariate normal

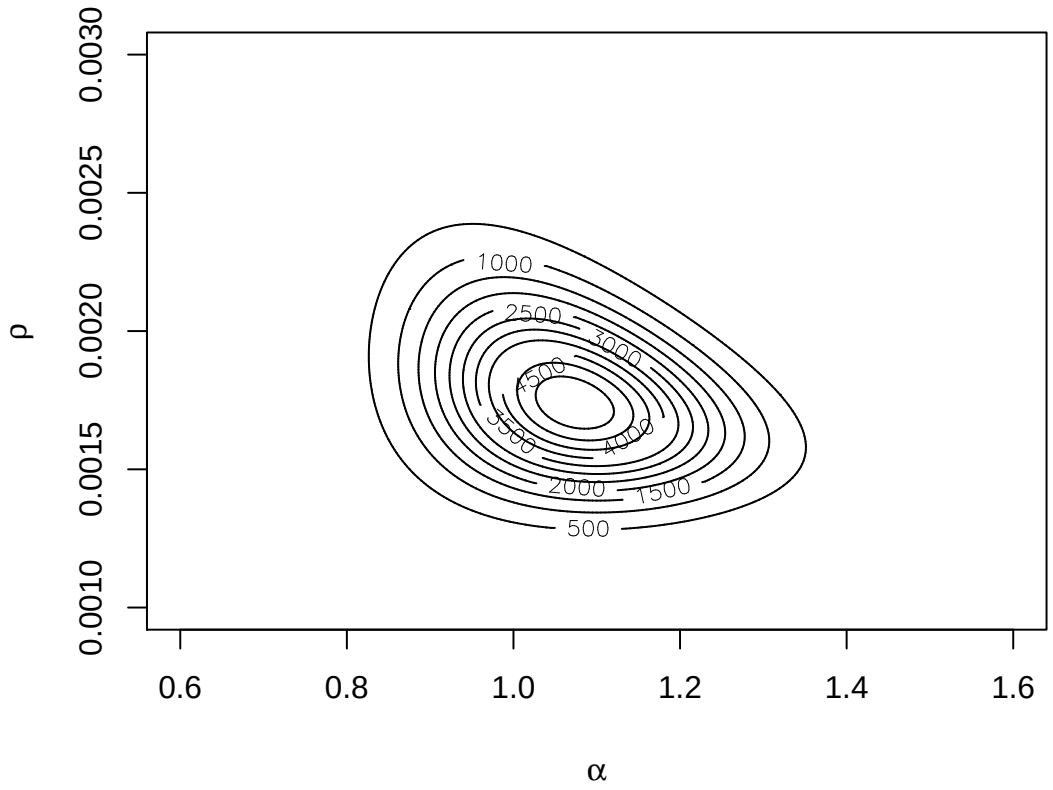

Figure 1: Posterior density of two unknowns: Contour plot

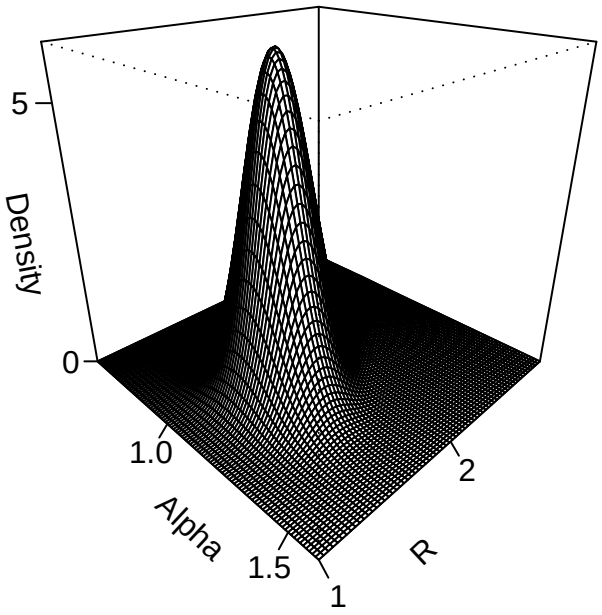

Figure 2: Posterior density of two unknowns: Wireframe plot

distribution with *mean vector* $\underline{\mu} = (\mu_1,\ \mu_2)^T$ and *variance matrix*

$$V = \begin{pmatrix} v_{1,1} & v_{1,2} \\ v_{1,2} & v_{2,2} \end{pmatrix}$$

where μ_1 and μ_2 are the means of Y_1 and Y_2 respectively, $v_{1,1}$ and $v_{2,2}$ are their variances, $v_{1,2}$ is their covariance and $|V|$ is the determinant of V.

If Y_1 and Y_2 are independent then $v_{1,2} = 0$ and, in the case of the bivariate normal distribution, the converse *is* true.

Note that, if X and Y both have normal marginal distributions it does not necessarily follow that their joint distribution is bivariate normal, although, in practice, the joint distribution often is bivariate normal. However, if X and Y both have normal distributions and are independent then their joint distribution is bivariate normal with zero covariance.

If Y_1 and Y_2 have a bivariate normal distribution then $a_1Y_1 + a_2Y_2$ is also normally distributed, where a_1 and a_2 are constants. For example, if $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ and X and Y are independent then $X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

0.1.3 Functions of continuous random variables (Revision)

Theory

As we shall see in the example below, we sometimes need to find the distribution of a random variable which is a function of another random variable. Suppose we have two random variables X and Y where $Y = g(X)$ for some function $g()$. In this section we will only consider the case where $g()$ is a strictly monotonic, i.e. either strictly increasing or strictly decreasing, function.

Suppose first that $g()$ is a strictly increasing function so that if $x_2 > x_1$ then $y_2 = g(x_2) > y_1 = g(x_1)$. In this case the distribution functions $F_X(x)$ and $F_Y(y)$ are related by

$$F_Y(y) = \Pr(Y < y) = \Pr(X < x) = F_X(x).$$

We can find the relationship between the probability density functions , $f_Y(y)$ and $f_X(x)$, by differentiating with respect to y. So

$$f_Y(y) = \frac{d}{dy}F_y(y) = \frac{d}{dy}F_X(x) = \frac{d}{dx}F_X(x) \times \frac{dx}{dy} = f_X(x)\frac{dx}{dy} = f_X(x)\left(\frac{dy}{dx}\right)^{-1}.$$

Similarly, if $g()$ is a strictly decreasing function so that if $x_2 > x_1$ then $y_2 = g(x_2) < y_1 = g(x_1)$,

$$F_Y(y) = \Pr(Y < y) = \Pr(X > x) = 1 - F_X(x)$$

and

$$f_Y(y) = -f_X(x)\frac{dx}{dy}$$

but here, of course, dx/dy is negative.

So, if $g()$ is a strictly monotonic function

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| \text{ where } \left|\frac{dx}{dy}\right| \text{ is the modulus of } \frac{dx}{dy}.$$

A simple way to remember this is to remember that an element of probability $f_X(x)\delta x$ is preserved through the transformation so that (for a strictly increasing function)

$$f_Y(y)\delta y = f_X(x)\delta x.$$

Example

Suppose for example that $X \sim N(\mu,\ \sigma^2)$ and that $Y = e^X$. So $X = \ln(Y)$ and $dx/dy = y^{-1}$. Now

$$f_X(x) = (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad (-\infty < x < \infty)$$

so

$$f_Y(y) = \frac{1}{y}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{\ln(y)-\mu}{\sigma}\right)^2\right\} \qquad (0 < y < \infty).$$

The resulting distribution for Y is called a *lognormal* distribution because $\ln(Y)$ has a normal distribution. It can be useful for representing beliefs about quantities which can only take positive values.

0.1.4 The multivariate normal distribution

Suppose that $\underline{X}$ has a multivariate normal $N_n(\underline{M},\ V)$ distribution. This distribution has a *mean vector* $\underline{M} = (m_1, \ldots, m_n)^T$ where m_i is the mean of X_i, and a *covariance matrix* V. The diagonal elements of V are the variances of $X_1, \ldots, X_n$ with the element in row and column i, v_{ii} being the variance of X_i. The covariance of X_i and X_j is v_{ij}, the element in row i and column j. Clearly $v_{ji} = v_{ij}$ and V is symmetric. It is also positive semi-definite.

The pdf is

$$f_X(\underline{x}) = (2\pi)^{-n/2}|V|^{-1/2}\exp\left\{-\frac{1}{2}(\underline{x}-\underline{M})^T V^{-1}(\underline{x}-\underline{M})\right\}.$$

(Here $\underline{x}^T$ denotes the transpose of $\underline{x}$). We often work in terms of the *precision matrix* $P = V^{-1}$. In this case, of course, we replace $(\underline{x}-\underline{M})^T V^{-1}(\underline{x}-\underline{M})$ with $(\underline{x}-\underline{M})^T P(\underline{x}-\underline{M})$.

If $\underline{X}$ has a multivariate normal $N_n(\underline{M},\ V)$ distribution and V is a diagonal matrix, that is if $\text{covar}(X_i, X_j) = 0$ when $i \neq j$, then $X_1, \ldots, X_n$ are independent.

0.1.5 Numerical Methods for More Than One Parameter

It is often necessary to use numerical methods to do the necessary integrations for computing posterior distributions and summaries. Such methods can be used when we have more than one unknown. We will look at this first in the case of two unknown parameters.

If we have two unknown parameters $\theta_1,\ \theta_2$ then we often need to create a two-dimensional grid of values, containing every combination of $\theta_{1,1}, \ldots, \theta_{1,m_1}$ and $\theta_{2,1}, \ldots, \theta_{2,m_2}$, where $\theta_{j,1}, \ldots, \theta_{j,m_j}$ are a set of, usually equally spaced, values of θ_j. We therefore have $m_1 m_2$ points and two step sizes, $\delta\theta_1,\ \delta\theta_2$. Figure 3 shows such a grid diagramatically. Instead of a collection of two-dimensional rectangular columns standing on a one-dimensional line, we now have a collection of three-dimensional rectangular columns standing on a two-dimensional plane. The contours in figure 3 represent the function being integrated. The small circles represent the points at which the function is evaluated. The dashed lines represent the boundaries of the columns. Of course we would really have many more function evaluations placed much more closely together. Notice that some of the function evaluations are in regions where the value of the function is very small. It is inefficient to waste too many function evaluations in this way and some more sophisticated methods avoid doing this.

The approximate integral becomes

$$\int\int h(\theta_1,\theta_2)\ d\theta_1\ d\theta_2 \approx \sum_{j=1}^{m_1}\sum_{k=1}^{m_2} h(\theta_{1,j},\theta_{2,k})\delta\theta_1\delta\theta_2.$$

We can extend this to three or more dimensions but it becomes impractical when the number of dimensions is large. If we use a 100×100 grid in two dimensions this gives 10^4 function evaluations. If we use a $100\times100\times100$ grid in three dimensions this requires 10^6 evaluations and so on. Clearly the number of evaluations becomes prohibitively large quite quickly as the number of dimensions increases. In such cases we would usually use Markov chain Monte Carlo methods which are beyond the scope of this module.

It is sometimes possible to reduce the dimension of the numerical integral by integrating analytically with respect to one unknown.

0.1.6 Example: The Weibull distribution

Model

The *Weibull distribution* is often used as a distribution for *lifetimes*. We might be interested, for example, in the lengths of time that a machine or component runs before it fails, or the survival time of a patient after a serious operation. A number of different families of distributions are used for such lifetime variables. Of course they are all continuous distributions and only give positive probability density to positive values of the lifetime. The Weibull distribution is an important distribution of this type. We can think of it as a generalisation of the exponential distribution. The distribution function of an exponential distribution is $F(t) = 1 - \exp(-\lambda t)$. The distribution function of a Weibull distribution is

$$F(t) = 1 - \exp\left(-\lambda t^\alpha\right) \qquad (t \geq 0) \tag{1}$$

where the extra parameter $\alpha > 0$ is called a *shape parameter*. It is often convenient to write $\lambda = \rho^\alpha$ and then

$$F(t) = 1 - \exp\left(-[\rho t]^\alpha\right) \qquad (t \geq 0) \tag{2}$$

and $\rho > 0$ is a *scale parameter*.

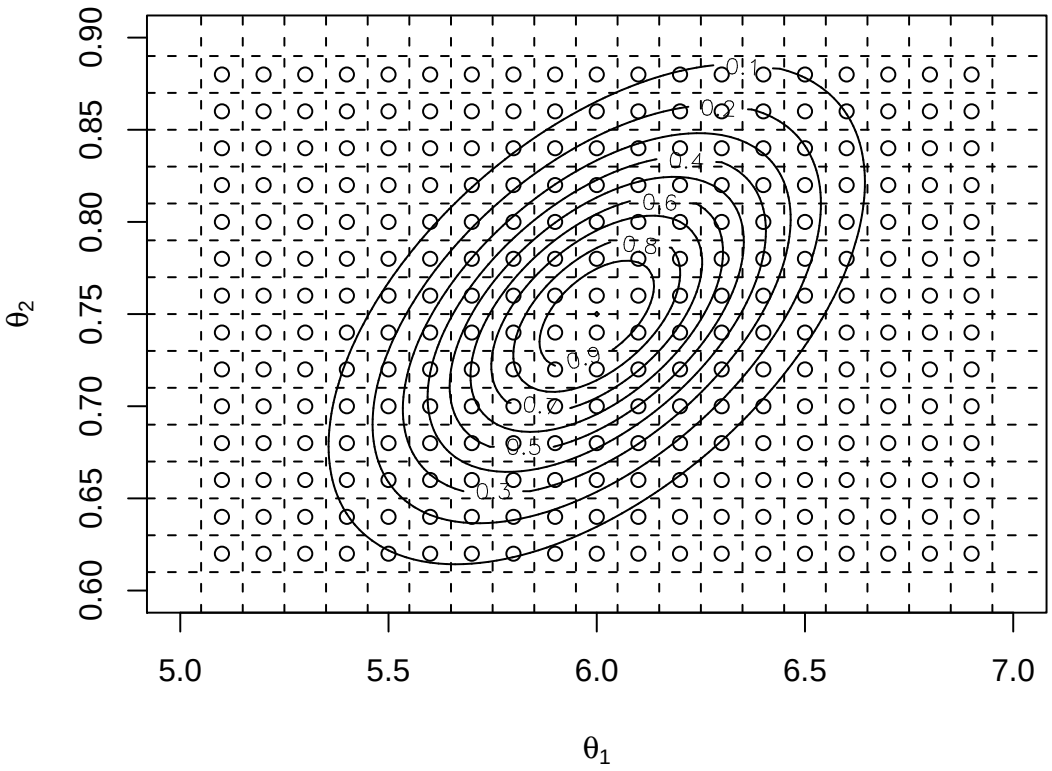

Figure 3: Numerical integration in two dimensions.

Differentiating (2) with respect to t, we obtain the pdf

$$f(t) = \alpha\rho(\rho t)^{\alpha-1}\exp\{-(\rho t)^{\alpha}\} \tag{3}$$

for $0 \le t < \infty$.

If we use $\alpha,\ \lambda$ instead of $\alpha,\ \rho$ as the parameters, as in (1), then the pdf is

$$f(t) = \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha}). \tag{4}$$

Evaluating the posterior distribution

Suppose that we work in terms of the $\alpha,\ \rho$ parameters of (3) and that we have n observations $t_1, \ldots, t_n$. Suppose that our prior density for α and ρ is $f^{(0)}_{\alpha,\rho}(\alpha,\rho)$. The likelihood is

$$L(\alpha,\rho) = \alpha^n\rho^{n\alpha}\left(\prod_{i=1}^{n} t_i\right)^{\alpha-1}\exp\left\{-\rho^{\alpha}\sum_{i=1}^{n} t_i^{\alpha}\right\}.$$

The posterior pdf is

$$f^{(1)}_{\alpha,\rho}(\alpha,\rho) \propto h_{\alpha,\rho}(\alpha,\rho) = f^{(0)}_{\alpha,\rho}(\alpha,\rho)L(\alpha,\rho).$$

To complete the evaluation of the posterior pdf we find

$$C = \int_0^{\infty}\int_0^{\infty} h_{\alpha,\rho}(\alpha,\rho)\ d\alpha\ d\rho$$

numerically and then

$$f^{(1)}_{\alpha,\rho}(\alpha,\rho) = h_{\alpha,\rho}(\alpha,\rho)/C.$$

Suppose, for example, that we give α and ρ independent gamma prior distributions so that

$$f^{(0)}_{\alpha,\rho}(\alpha,\rho) \propto \alpha^{a_\alpha-1}e^{-b_\alpha\alpha}\rho^{a_\rho-1}e^{-b_\rho\rho}.$$

Then the posterior pdf is proportional to

$$h_{\alpha,\rho}(\alpha,\rho) = \alpha^{n+a_\alpha-1}\rho^{n\alpha+a_\rho-1}\left(\prod_{i=1}^{n} t_i\right)^{\alpha-1}\exp\left\{-\left[b_\alpha\alpha + b_\rho\rho + \rho^{\alpha}\sum_{i=1}^{n} t_i^{\alpha}\right]\right\}.$$

Figure 1 shows the posterior density of α and ρ when $n = 50,\ a_\alpha = 1,\ b_\alpha = 1,\ a_\rho = 3,\ b_\rho = 1000$ and the data are as given in table 1. Figure 2 shows the same thing as a perspective plot except that, to make the axes more readable, ρ has been replaced with $R = 1000\rho$.

To find, for example, the posterior mean of ρ we evaluate

$$\int_0^{\infty}\int_0^{\infty} \rho\ f^{(1)}_{\alpha,\rho}(\alpha,\rho)\ d\alpha\ d\rho = C^{-1}\int_0^{\infty}\int_0^{\infty} \rho\ h_{\alpha,\rho}(\alpha,\rho)\ d\alpha\ d\rho.$$

To find a 95 % hpd region for α, ρ we can either choose a value k and evaluate $\int\int f^{(1)}_{\alpha,\rho}(\alpha,\rho)\ d\alpha\ d\rho$ over all points in a grid for which $f^{(1)}_{\alpha,\rho}(\alpha,\rho) > k$ then adjust k and repeat until the value of 0.95

67	313	1391	630	627	573	2093	28	492	482
206	1166	165	1088	496	313	437	815	436	17
32	131	340	939	247	1859	57	132	813	254
950	1615	463	258	2285	672	506	50	637	246
178	431	306	662	33	254	858	187	344	545

Table 1: Data for Weibull example.

is obtained or rank all of the points in our grid in decreasing order of $f^{(1)}_{\alpha,\rho}(\alpha,\rho)$ and cumulatively integrate over them until 0.95 is reached.

To find the marginal pdf for α we evaluate

$$\int_0^\infty f^{(1)}_{\alpha,\rho}(\alpha,\rho)\ d\rho.$$

0.1.7 Transformations

Theory

It has probably become apparent by now that sometimes it may be helpful to use a transformation of the parameters. For example, sometimes a posterior distribution where we need to use numerical integration might have an awkward shape which makes placing a suitable and efficient rectangular grid difficult.

In section 0.1.3 we saw how to change the pdf when we transform a single random variable. Sometimes, of course, we need a more general method for transforming between one set of parameters and another. Let $\underline{\theta}$ and $\underline{\phi}$ be two alternative sets of parameters where there is a 1 - 1 relationship between values of $\underline{\theta}$ and values of $\underline{\phi}$, and therefore each contains the same number of parameters. (There could appear to be more parameters in $\underline{\theta}$ than in $\underline{\phi}$, for example, but, in that case, there would have to be constraints on the values of $\underline{\theta}$ so that there was the same *effective* number of parameters in $\underline{\theta}$ and $\underline{\phi}$). Let $\underline{\theta} = (\theta_1, \ldots, \theta_k)^T$ and $\underline{\phi} = (\phi_1, \ldots, \phi_k)^T$. Suppose also that we can write, for each i,

$$\phi_i = g_i(\theta_1, \ldots, \theta_k)$$

where g is a differentiable function. Then, if the density of $\underline{\theta}$ is $f_{\underline{\theta}}(\underline{\theta})$ and the density of $\underline{\phi}$ is $f_{\underline{\phi}}(\underline{\phi})$,

$$f_{\underline{\theta}}(\underline{\theta}) = f_{\underline{\phi}}(\underline{\phi})|J|$$

where J is the *Jacobian determinant*, often just called "the Jacobian,"

$$\begin{vmatrix} \frac{\partial\phi_1}{\partial\theta_1} & \frac{\partial\phi_1}{\partial\theta_2} & \cdots & \frac{\partial\phi_1}{\partial\theta_k} \\ \frac{\partial\phi_2}{\partial\theta_1} & \frac{\partial\phi_2}{\partial\theta_2} & \cdots & \frac{\partial\phi_2}{\partial\theta_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial\phi_k}{\partial\theta_1} & \frac{\partial\phi_k}{\partial\theta_2} & \cdots & \frac{\partial\phi_k}{\partial\theta_k} \end{vmatrix}$$

and $|J|$ is its modulus.

For example, we could transform the $(0,\infty)$ ranges of the parameters α, ρ of a Weibull distribution to $(0,1)$ by using

$$\beta = \frac{\alpha}{\alpha+1}, \qquad \gamma = \frac{\rho}{\rho+1}.$$

The Jacobian is

$$J = \begin{vmatrix} \frac{\partial\beta}{\partial\alpha} & \frac{\partial\beta}{\partial\rho} \\ \frac{\partial\gamma}{\partial\alpha} & \frac{\partial\gamma}{\partial\rho} \end{vmatrix} = (\alpha+1)^{-2}(\rho+1)^{-2}.$$

Suppose that the joint posterior density of α and ρ is proportional to $h_{\alpha,\rho}(\alpha,\rho)$. So we define

$$h_{\beta,\gamma}(\beta,\gamma) = (\alpha+1)^2(\rho+1)^2 h_{\alpha,\rho}(\alpha,\rho),$$

where

$$\alpha = \frac{\beta}{1-\beta}, \qquad \rho = \frac{\gamma}{1-\gamma}$$

so

$$h_{\beta,\gamma}(\beta,\gamma) = (1-\beta)^{-2}(1-\rho)^{-2} h_{\alpha,\rho}\left(\frac{\beta}{1-\beta}, \frac{\gamma}{1-\gamma}\right).$$

Then let

$$C = \int_0^1 \int_0^1 h_{\beta,\gamma}(\beta,\gamma)\ d\beta\ d\gamma.$$

The posterior mean of ρ is then

$$C^{-1} \int_0^1 \int_0^1 \frac{\gamma}{1-\gamma} h_{\beta,\gamma}(\beta,\gamma)\ d\beta\ d\gamma.$$

A hpd region for α, ρ can then be found by integrating $C^{-1}h_{\beta,\gamma}(\beta,\gamma)$ with respect to β, γ over the points with the greatest values of

$$h_{\alpha,\rho}\left(\frac{\beta}{1-\beta}, \frac{\gamma}{1-\gamma}\right) = h_{\alpha,\rho}(\alpha,\rho).$$

Example: A clinical trial

The Anturane Reinfarction Trial Research Group (1980) reported a clinical trial on the use of the drug sulfinpyrazone in patients who had suffered myocardial infarctions ("heart attacks"). The idea was to see whether the drug had an effect on the number dying. Patients in one group were given the drug while patients in another group were given a "placebo," that is an inactive substitute. The following table gives the number of all "analysable" deaths up to 24 months after the myocardial infarction and the total number of eligible patients who were not withdrawn and did not suffer a "non-analysable" death during the study.

	Deaths	Total
Group 1 (Sulfinpyrazone)	44	560
Group 2 (Placebo)	62	540

We can represent this situation by saying that there are two groups, containing n_1 and n_2 patients, and two parameters, $\theta_1,\ \theta_2$, such that, given these parameters, the distribution of the number of deaths X_j in Group j is binomial$(n_j,\ \theta_j)$.

Now we could give θ_j a beta prior distribution but it seems reasonable that our prior beliefs would be such that θ_1 and θ_2 would not be independent. There are various ways in which we could represent this. One of these is as follows. We transform from the $(0,1)$ scale of $\theta_1,\ \theta_2$ to a $(-\infty,\infty)$ scale and then give the new parameters, $\eta_1,\ \eta_2$, a bivariate normal distribution (see section 0.1.2). We can use a transformation where $\theta_j = F(\eta_j)$ and $F(x)$ is the distribution function of a continuous distribution on $(-\infty,\ \infty)$, usually one which is symmetric about $x = 0$. One possibility is to use the standard normal distribution function $\Phi(x)$ so that $\theta_j = \Phi(\eta_j)$. We write $\eta_j = \Phi^{-1}(\theta_j)$ where this function, $\Phi^{-1}(x)$, the inverse of the standard normal distribution function, is sometimes called the *probit* function. If we use this transformation then it is easily seen that

$$f_\theta(\theta_1,\theta_2) = f_\eta(\eta_1,\eta_2)/|J|,$$

where $f_\theta(\theta_1,\theta_2)$ is the joint density of $\theta_1,\ \theta_2$, $f_\eta(\eta_1,\ \eta_2)$ is the joint density of $\eta_1,\ \eta_2$ and

$$|J| = \left| \begin{array}{cc} \frac{\partial\theta_1}{\partial\eta_1} & \frac{\partial\theta_1}{\partial\eta_2} \\ \frac{\partial\theta_2}{\partial\eta_1} & \frac{\partial\theta_2}{\partial\eta_2} \end{array} \right| = \phi(\eta_1)\phi(\eta_2),$$

where $\phi(x)$ is the standard normal pdf.

Suppose that, from past experience, we can give a 95% symmetric prior interval for θ_2 (placebo) as $0.05 < \theta_2 < 0.20$. (This is actually quite a wide interval considering that there may be a lot of past experience of such patients). This converts to a 95% interval of $-1.645 < \eta_2 < -0.842$. For example, in R, we can use

```
> qnorm(0.025,0,1)
[1] -1.959964
```

If we give η_2 a normal prior distribution then we require the mean to be $\mu_2 = ([-1.645] + [-0.842])/2 \approx -1.24$ and the standard deviation to be $\sigma_2 = ([-0.842]-[-1.645])/(2\times 1.96) \approx 0.21$, since a symmetric 95% normal interval is the mean plus or minus 1.96 standard deviations. Let us use the same mean for a normal prior distribution for η_1 (sulfinpyrazone) so that we have equal prior probabilities for an increase and a decrease in death rate when the treatment is given. However it seems reasonable that we would be less certain of the death rate given the treatment so we increase the prior standard deviation to $\sigma_1 = 2\sigma_2 = 0.42$. This implies a 95% interval $-2.06 < \eta_1 < -0.42$ which, in turn, implies $0.02 < \theta_1 < 0.34$. (This is a wide interval so we are really not supplying much prior information).

We also need to choose a covariance or correlation between η_1 and η_2. At this point we will not discuss in detail how to do this except to say that, if we choose the correlation to be r, then the conditional variance of one of $\eta_1,\ \eta_2$ given the other will be $100r^2\%$ of the marginal variance. For example, if we choose $r = 0.7$, then the variance of one is roughly halved by learning the value of the other. Suppose that we choose this value. Then the covariance between η_1 and η_2 is $0.7 \times 0.21 \times 0.42 = 0.0617$.

In evaluating the joint prior density of $\eta_1,\ \eta_2$, we can make use of the fact, which is easily confirmed, that, if $\delta_j = (\eta_j - \mu_j)/\sigma_j$ and $r = \text{covar}(\eta_1, \eta_2)/(\sigma_1\sigma_2)$, then the joint density is proportional to

$$\exp\left\{-\frac{1}{2(1-r^2)}(\delta_1^2 + \delta_2^2 - 2r\delta_1\delta_2)\right\}.$$

Figure 4 shows a R function to evaluate the posterior density. Figure 5 shows the resulting posterior density. The dashed line is the line $\theta_1 = \theta_2$. We see that most of the probability lies on the side where $\theta_2 > \theta_1$ which suggests that the death rate is probably greater with the placebo than with sulfinpyrazone, which, of course, suggests that sulfinpyrazone has a beneficial effect.

To investigate further what the posterior tells us about the effect of sulfinpyrazone, we can calculate the posterior probability that $\theta_1 < \theta_2$. This is done by integrating the joint posterior density over the region where $\theta_1 < \theta_2$. This calculation is included in the function shown in figure 4. The calculated probability is 0.972. We can also find the posterior density of the *relative risk*, θ_1/θ_2, or the *log relative risk*, $\log(\theta_1/\theta_2)$. Let γ be the log relative risk. We can modify the function in figure 4 so that it uses a grid of γ and θ_2 values, evaluates the joint posterior density of γ and θ_2 and then integrates out θ_2. Of course we need to transform between $\theta_1,\ \theta_2$ and $\gamma,\ \theta_2$ where the densities are related by

$$f_{\theta_1,\theta_2}(\theta_1, \theta_2) = f_{\gamma,\theta_2}(\gamma, \theta_2)|J|$$

and $J = \theta_1^{-1}$ is the appropriate Jacobian. Figure 6 shows the prior and posterior densities of the log relative risk, γ. Values of γ less than zero correspond to a smaller death rate with sulfinpyrazone than with the placebo. Notice that the prior density is not quite symmetric about zero. It is symmetric on the η scale but not on the γ scale. The prior median is zero, however.

There are other methods available to deal with problems of this sort, some involving approximations and fairly simple calculations.

```
function(theta1,theta2,n,x,prior)
{# Evaluates posterior density for probit example.
 # prior is mean1, mean2, sd1, sd2, correlation
 n1<-length(theta1)
 n2<-length(theta2)
 step1<-theta1[2]-theta1[1]
 step2<-theta2[2]-theta2[1]
 theta1<-matrix(theta1,nrow=n1,ncol=n2)
 theta2<-matrix(theta2,nrow=n1,ncol=n2,byrow=T)
 eta1<-qnorm(theta1,0,1)
 eta2<-qnorm(theta2,0,1)
 delta1<-(eta1-prior[1])/prior[3]
 delta2<-(eta2-prior[2])/prior[4]
 r<-prior[5]
 d<-1-r^2
 logprior<- -(delta1^2 + delta2^2 - 2*r*delta1*delta2)/(2*d)
 J<-dnorm(eta1,0,1)*dnorm(eta2,0,1)
 logprior<-logprior-log(J)
 loglik<-x[1]*log(theta1)+(n[1]-x[1])*log(1-theta1)+x[2]*log(theta2)+(n[2]-x[2])*log(1-theta2)
 logpos<-logprior+loglik
 logpos<-logpos-max(logpos)
 posterior<-exp(logpos)
 int<-sum(posterior)*step1*step2
 posterior<-posterior/int
 prob<-sum(posterior*(theta1<theta2))*step1*step2
 ans<-list(density=posterior,prob=prob)
 ans
 }
```

Figure 4: R function for probit example (0.1.7).

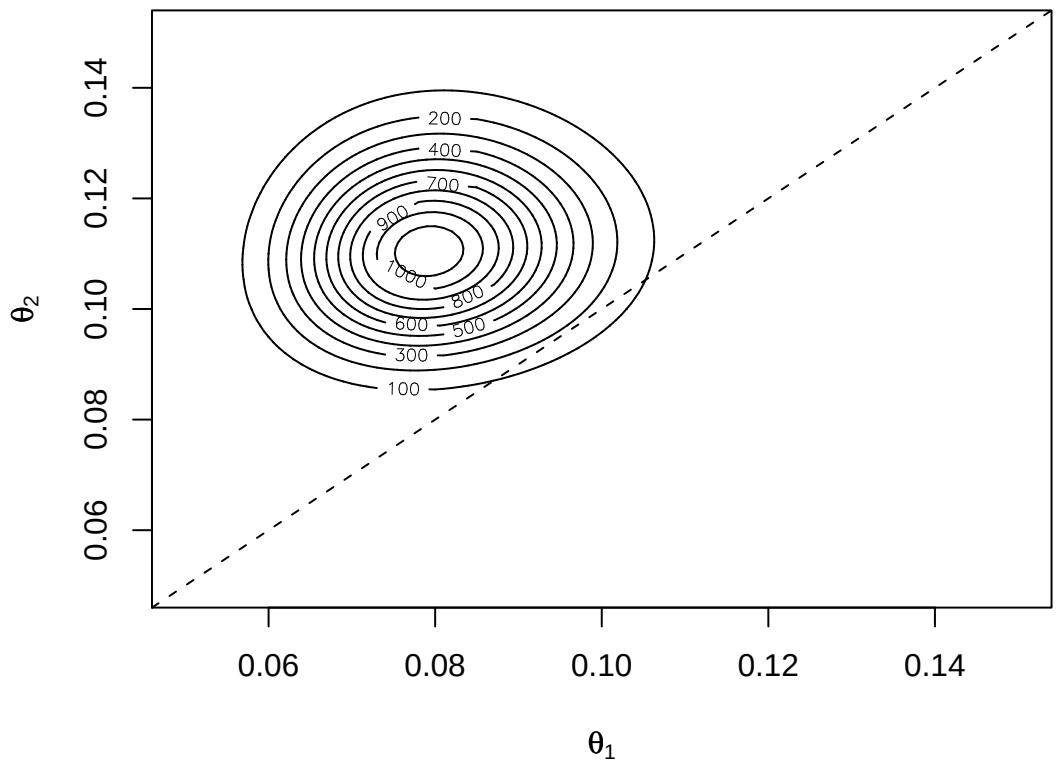

Figure 5: Posterior density of θ_1 and θ_2 in probit example (0.1.7). The dashed line is $\theta_1 = \theta_2$.

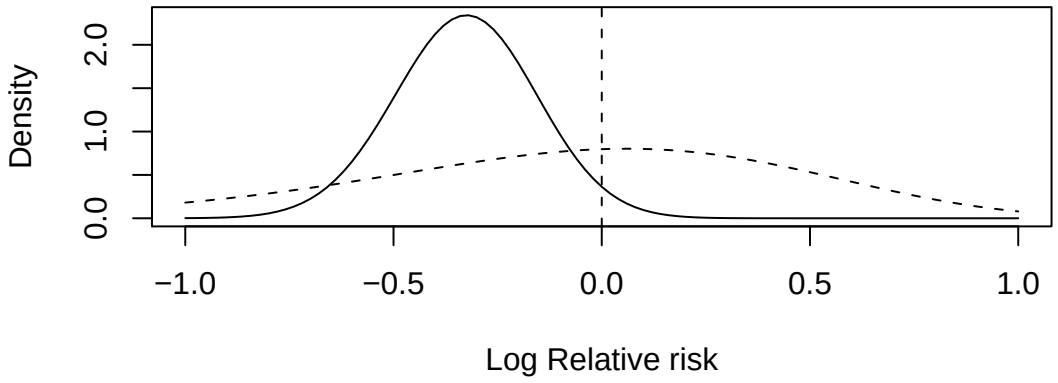

Figure 6: Posterior density (solid) and prior density (dashes) of log relative risk in probit example (0.1.7).

0.2 The Dirichlet distribution and multinomial observations

0.2.1 The Dirichlet distribution

The Dirichlet distribution is a distribution for a set of quantities $\theta_1, \ldots, \theta_m$ where $\theta_i \geq 0$ and $\sum_{i=1}^m \theta_i = 1$. An obvious application is to a set of probabilities for a partition (i.e. for an exhaustive set of mutually exclusive events).

The probability density function is

$$f(\theta_1, \ldots, \theta_m) = \frac{\Gamma(A)}{\prod_{i=1}^m \Gamma(a_i)} \prod_{i=1}^m \theta_i^{a_i - 1}$$

where $A = \sum_{i=1}^m a_i$ and $a_1, \ldots, a_m$ are parameters with $a_i > 0$ for $i = 1, \ldots, m$. We write $\mathrm{D}_m(a_1, \ldots, a_m)$ for this distribution.

Clearly, if $m = 2$, we obtain a $\mathrm{Beta}(a_1, a_2)$ distribution as a special case.

The mean of θ_j is

$$\mathrm{E}(\theta_j) = \frac{a_j}{A}$$

the variance of θ_j is

$$\mathrm{var}(\theta_j) = \frac{a_j}{A(A+1)} - \frac{a_j^2}{A^2(A+1)}$$

and the covariance of θ_j and θ_k, where $j \neq k$, is

$$\mathrm{covar}(\theta_j, \theta_k) = -\frac{a_j a_k}{A^2(A+1)}.$$

Also the marginal distribution of θ_j is $\mathrm{Beta}(a_j,\ A - a_j)$.

Note that the space of the parameters $\theta_1, \ldots, \theta_m$ has only $m - 1$ dimensions because of the constraint $\sum_{i=1}^m \theta_i = 1$, so that, for example, $\theta_m = 1 - \sum_{i=1}^{m-1} \theta_i$. Therefore, when we integrate over this space, the integration has only $m - 1$ dimensions.

Proof (mean)

The mean is

$$\begin{aligned}
\mathrm{E}(\theta_j) &= \int \cdots \int \theta_j \frac{\Gamma(A)}{\prod_{i=1}^m \Gamma(a_i)} \prod_{i=1}^m \theta_i^{a_i - 1}\ d\theta_1 \ldots d\theta_{m-1} \\
&= \frac{\Gamma(A)}{\Gamma(A+1)} \frac{\Gamma(a_j + 1)}{\Gamma(a_j)} \int \cdots \int \frac{\Gamma(A+1)}{\prod_{i=1}^m \Gamma(a_i')} \prod_{i=1}^m \theta_i^{a_i' - 1}\ d\theta_1 \ldots d\theta_{m-1} \\
&= \frac{\Gamma(A)}{\Gamma(A+1)} \frac{\Gamma(a_j + 1)}{\Gamma(a_j)} = \frac{a_j}{A}
\end{aligned}$$

where $a_i' = a_i$ when $i \neq j$ and $a_j' = a_j + 1$.

Proof (variance)

Similarly

$$\mathrm{E}(\theta_j^2) = \frac{\Gamma(A)}{\Gamma(A+2)}\frac{\Gamma(a_j+2)}{\Gamma(a_j)} = \frac{(a_j+1)a_j}{(A+1)A}$$

so

$$\mathrm{var}(\theta_j) = \frac{(a_j+1)a_j}{(A+1)A} - \Big(\frac{a_j}{A}\Big)^2 = \frac{a_j}{A(A+1)} - \frac{a_j^2}{A^2(A+1)}$$

Proof (covariance)

Also

$$\mathrm{E}(\theta_j\theta_k) = \frac{\Gamma(A)}{\Gamma(A+2)}\frac{\Gamma(a_j+1)}{\Gamma(a_j)}\frac{\Gamma(a_k+1)}{\Gamma(a_k)} = \frac{a_j a_k}{(A+1)A}$$

so

$$\mathrm{covar}(\theta_j,\theta_k) = \frac{a_j a_k}{(A+1)A} - \frac{a_j}{A}\frac{a_k}{A} = -\frac{a_j a_k}{A^2(A+1)}$$

Proof (marginal)

We can write the joint density of $\theta_1,\ldots,\theta_m$ as

$$f_1(\theta_1)f_2(\theta_2 \mid \theta_1)f_3(\theta_3 \mid \theta_1,\theta_2)\cdots f_{m-1}(\theta_{m-1} \mid \theta_1,\ldots,\theta_{m-2}).$$

(We do not need to include a final term in this for θ_m because θ_m is fixed once $\theta_1,\ldots,\theta_{m-1}$ are fixed).

In fact we can write the joint density as

$$\frac{\Gamma(A)}{\Gamma(a_1)\Gamma(A-a_1)}\theta_1^{a_1-1}(1-\theta_1)^{A-a_1-1} \times \frac{\Gamma(A-a_1)}{\Gamma(a_2)\Gamma(A-a_1-a_2)}\frac{\theta_2^{a_2-1}(1-\theta_1-\theta_2)^{A-a_1-a_2-1}}{(1-\theta_1)^{A-a_1-1}}$$

$$\times\cdots\times \frac{\Gamma(A-a_1-\cdots-a_{m-2})}{\Gamma(a_{m-1})\Gamma(A-a_1-\cdots-a_{m-1})}\frac{\theta_{m-1}^{a_{m-1}-1}\theta_m^{a_m-1}}{(1-\theta_1-\cdots\theta_{m-2})^{a_{m-1}+a_m-1}}.$$

A bit of cancelling shows that this simplifies to the correct Dirichlet density.

Thus we can see that the marginal distribution of θ_1 is a Beta$(a_1,\ A-a_1)$ distribution and similarly that the marginal distribution of θ_j is a Beta$(a_j,\ A-a_j)$ distribution. We can also deduce the distribution of a subset of $\theta_1,\ldots,\theta_m$. For example if $\tilde{\theta}_3 = 1-\theta_1-\theta_2-\theta_3$, then the distribution of $\theta_1,\theta_2,\theta_3,\tilde{\theta}_3$ is Dirichlet $\mathrm{D}_d(a_1,\ a_2,\ a_3,\ \tilde{a}_3)$ where $\tilde{a}_3 = A-a_1-a_2-a_3$.

0.2.2 Multinomial observations

Model

Suppose that we will observe $X_1,\ldots,X_m$ where these are the frequencies for categories $1,\ldots,m$, the total $N=\sum_{i=1}^m X_i$ is fixed and the probabilities for these categories are $\theta_1,\ldots,\theta_m$ where $\sum_{i=1}^m \theta_i = 1$. Then, given θ, where $\theta=(\theta_1,\ldots,\theta_m)^T$, the distribution of $X_1,\ldots,X_m$ is multinomial with

$$\Pr(X_1=x_1,\ldots,X_m=x_m) = \frac{N!}{\prod_{i=1}^m x_i!}\prod_{i=1}^m \theta_i^{x_i}.$$

Notice that, with $m=2$, this is just a Bin(N,θ_1) distribution.

Then the likelihood is

$$\begin{aligned} L(\theta;\ x) &= \frac{N!}{\prod_{i=1}^m x_i!}\prod_{i=1}^m \theta_i^{x_i} \\ &\propto \prod_{i=1}^m \theta_i^{x_i}. \end{aligned}$$

The conjugate prior is a *Dirichlet* distribution which has a pdf proportional to

$$\prod_{i=1}^m \theta_i^{a_i-1}.$$

The posterior pdf is proportional to

$$\prod_{i=1}^m \theta_i^{a_i-1} \times \prod_{i=1}^m \theta_i^{x_i} = \prod_{i=1}^m \theta_i^{a_i+x_i-1}.$$

This is proportional to the pdf of a Dirichlet distribution with parameters $a_1+x_1, a_2+x_2, \ldots a_m+x_m$.

Example

In a survey 1000 English voters are asked to say for which party they would vote if there were a general election next week. The choices offered were 1: Labour, 2: Liberal, 3: Conservative, 4: Other, 5: None, 6: Undecided. We assume that the population is large enough so that the responses may be considered independent given the true underlying proportions. Let $\theta_1,\ldots,\theta_6$ be the probabilities that a randomly selected voter would give each of the responses. Our prior distribution for $\theta_1,\ldots,\theta_6$ is a $\mathrm{D}_6(5,\ 3,\ 5,\ 1,\ 2,\ 4)$ distribution.

This gives the following summary of the prior distribution.

Response	a_i	Prior mean	Prior var.	Prior sd.
Labour	5	0.25	0.008929	0.09449
Liberal	3	0.15	0.006071	0.07792
Conservative	5	0.25	0.008929	0.09449
Other	1	0.05	0.002262	0.04756
None	2	0.10	0.004286	0.06547
Undecided	4	0.20	0.007619	0.08729
Total	20	1.00		

Suppose our observed data are as follows.

Labour	Liberal	Conservative	Other	None	Undecided
256	131	266	38	114	195

Then we can summarise the posterior distribution as follows.

Response	$a_i + x_i$	Posterior mean	Posterior var.	Posterior sd.
Labour	261	0.2559	0.0001865	0.01366
Liberal	134	0.1314	0.0001118	0.01057
Conservative	271	0.2657	0.0001911	0.01382
Other	39	0.0382	0.0000360	0.00600
None	116	0.1137	0.0000987	0.00994
Undecided	199	0.1951	0.0001538	0.01240
Total	1020	1.0000		

Chapter 1

The Normal Linear Model

1.1 Regression and the normal linear model

1.1.1 Introduction

A model which describes how the conditional distribution of one variable, often called the *dependent* variable, given some other variables, depends on the values taken by these other variables, is called a *regression*. Typically we are interested in how the conditional mean of the dependent variable depends on the values of the other variables but other features of the distribution may also change. Various names are used to describe these other variables, including *regressors*, *explanatory variables* and *covariates*.

There are many different kinds of regression models. One of the simplest is described by the equation

$$Y = \alpha + \beta x + \varepsilon. \tag{1.1}$$

This might be used in situations where an observation consists of a pair of values $(x_i,\ y_i)$ where

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

- y_i is observation number i on the dependent variable,
- x_i is observation number i on a single explanatory variable,
- ε_i is a random "error" and
- α and β are parameters the values of which are usually unknown.

Our data might consist of n such pairs.

We also need to specify a sampling distribution for ε_i. In this chapter we assume the following *(conditional on model parameters)*:

Normality : $\varepsilon \sim N(0,\ \sigma^2)$.

Independence : $\varepsilon_1, \ldots, \varepsilon_n$ are independent , given the parameters of their distribution (typically the variance σ^2).

Equality of variance each of $\varepsilon_1, \ldots, \varepsilon_n$ has the same variance σ^2 (equivalently, the same precision τ).

Another way to express this model is to say that the conditional distribution of Y given x (and the model parameters) is normal with mean $\alpha + \beta x$ and variance σ^2 and that, given x_i and x_j and the model parameters, Y_i and Y_j are independent for $i \neq j$.

This model, with the relationship given in (1.1) and these assumptions about the errors is called an ordinary linear regression on a single covariate with normal errors.

1.1.2 Example

Here is a simple example. We wish to be able to predict the height of a student if we know the student's shoe size.

Suppose that we are prepared to accept (1.1) as a reasonable description of the relationship. That is, the conditional mean height, given shoe size, is a linear function of shoe size and the actual heights, given a particular shoe size, have a distribution centered on this mean. The "errors" ε are the differences between the actual height values and the conditional mean given by our linear function of shoe size. Suppose that we are also prepared to accept the usual assumptions of normality, independence and equal variance, that is that the conditional variance of height given shoe size does not depend on shoe size. These are issues of *model choice*. One way to think of a regression model like this is as a device which allows us to use information from many different values of the regressor X to help us to make predictions about Y for other values of X, in a way which seems to be appropriate to us according to our prior beliefs.

We need to specify our prior distribution for the parameters. There are three parameters in this model, α, the intercept, β, the slope of the *regression line*, and $\tau = \sigma^{-2}$, the error precision.

There are many possibilities, including the following.

- The value of τ is known and we give α and β a bivariate normal prior.
- The value of τ is unknown but we use a conjugate prior. We give τ a gamma prior and we give α and β a bivariate normal conditional prior, given τ, where the precision matrix is proportional to τ.
- We use a semi-conjugate prior in which τ has a gamma prior and α and β have a bivariate normal prior independently of τ.
- A non-conjugate prior.

For illustration, consider a semi-conjugate prior.

As it stands, our model says that the conditional mean height for a student with shoe size x is $\alpha + \beta x$. This makes α a rather unnatural paramater because it represents the mean height for students with shoe size zero, a shoe size which is well outside the usual range for students. This makes it difficult for us to think about our prior beliefs about α and also creates a rather awkward relationship between α and β in our beliefs since a change in α would require a change in β to make the regression line continue to pass through the region where we think (X, Y) points will typically be found. It is better to change the origin of X to a more usable reference value x_{ref}. I know my own shoe size and height so let us use my shoe size, 11, as a reference value. Let $z = x - x_{\text{ref1}} = x - 11$ then our regression equation becomes

$$Y = \tilde{\alpha}_1 + \beta z + \varepsilon,$$

where $\tilde{\alpha}_1 = \alpha + \beta x_{\text{ref1}} = \alpha + 11\beta$ now represents the mean height for students who take size 11 shoes. I can now use my own height, 74 inches, as a guide to the likely value of $\tilde{\alpha}$. Let us give $\tilde{\alpha}$ a prior distribution which is normal with mean 74. Of course I can not assume that I am exactly the average height for size 11 shoe-wearers so we need a suitable prior standard deviation for $\tilde{\alpha}_1$. I think that, even bearing in mind that it is a long time since I was a first year student, I am unlikely to be more than six inches from the conditional mean so let us make the standard deviation 3, giving a variance of 9 and a precision of 0.111. We could choose to make the standard deviation larger, of course, if we felt less confident about the value of our prior information.

Now we need a prior for β. How much does the mean height change when we change the shoe size by one unit? As a guide, my wife is 64 inches tall and takes size 5 shoes. This suggests a change of 10 inches in 6 shoes sizes or $\beta = 10/6 \approx 1.7$. Let us use $x_{\text{ref2}} = 5$ as a second reference value. Let $\tilde{\alpha}_2 = \alpha + \beta x_{\text{ref2}} = \alpha + 5\beta$ now represent the mean height for students who take size 5 shoes. Let us give $\tilde{\alpha}_2$ a $N(64,\ 9)$ prior distribution.

In this example it seems reasonable to make $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ independent in our prior distribution and this is what we will do. In other examples we might, for example, feel that we are likely to have misjudged both conditional means in the same direction and so give them a positive covariance. So, let us write

$$\underset{\sim}{\tilde{\beta}} = \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{pmatrix}.$$

Then $\underline{\tilde{\beta}}$ has a bivariate normal prior distribution with mean

$$\underline{\tilde{M}} = \begin{pmatrix} \mathrm{E}(\tilde{\alpha}_1) \\ \mathrm{E}(\tilde{\alpha}_2) \end{pmatrix} = \begin{pmatrix} 74 \\ 64 \end{pmatrix}$$

and variance

$$\tilde{V}_0 = \begin{pmatrix} \mathrm{var}(\tilde{\alpha}_1) & \mathrm{covar}(\tilde{\alpha}_1, \tilde{\alpha}_2) \\ \mathrm{covar}(\tilde{\alpha}_1, \tilde{\alpha}_2) & \mathrm{var}(\tilde{\alpha}_2) \end{pmatrix} = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}.$$

It is easily seen that

$$\beta = \frac{\tilde{\alpha}_1 - \tilde{\alpha}_2}{x_{\mathrm{ref}1} - x_{\mathrm{ref}2}} \qquad \text{and} \qquad \alpha = \frac{\tilde{\alpha}_2 x_{\mathrm{ref}1} - \tilde{\alpha}_1 x_{\mathrm{ref}2}}{x_{\mathrm{ref}1} - x_{\mathrm{ref}2}}.$$

We could continue to work with, for example, z rather than x but we can convert back to the original parameters using

$$\underline{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{x_{\mathrm{ref}1} - x_{\mathrm{ref}2}} \begin{pmatrix} -x_{\mathrm{ref}2} & x_{\mathrm{ref}1} \\ 1 & -1 \end{pmatrix} \underline{\tilde{\beta}} = H\underline{\tilde{\beta}}.$$

Then $\alpha,\ \beta$ have a bivariate normal prior distribution with mean

$$\underline{M}_0 = H\underline{\tilde{M}} = \frac{1}{6} \begin{pmatrix} -5 & 11 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 74 \\ 64 \end{pmatrix} = \begin{pmatrix} 55.7 \\ 1.7 \end{pmatrix}$$

and variance

$$V_0 = H\tilde{V}_0 H^T = \frac{1}{36} \begin{pmatrix} -5 & 11 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix} \begin{pmatrix} -5 & 1 \\ 11 & -1 \end{pmatrix} = \begin{pmatrix} 36.5 & -4 \\ -4 & 0.5 \end{pmatrix}.$$

1.1.3 The normal linear model

The normal linear model is a more general class of models which includes (1.1) and many more kinds of model, as special cases.

First of all let us rewrite (1.1), using slightly different notation, as

$$Y = \beta_0 + \beta_1 x + \varepsilon. \tag{1.2}$$

Now suppose that we want to relate the dependent variable Y to the values, $x_1, \ldots, x_k$, of two or more regressors $X_1, \ldots, X_k$. One way to do this is to write

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

So, our model for observation i is

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \varepsilon_i, \tag{1.3}$$

where $x_{i,j}$ is the value of regressor X_j in observation i.

It is convenient to make a further change to the notation. We relabel the regressors and coefficients $1, \ldots, p$ instead of $0, \ldots, k$. So $k = p - 1$. Then

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i = \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i. \tag{1.4}$$

We seem to have lost the intercept term β_0 in (1.2) and (1.3). However this is easily overcome by defining X_1 so that $x_{i,1} = 1$ for all i. Then we can rewrite (1.2) as

$$Y = \beta_1 1 + \beta_2 x + \varepsilon$$

and define $X_1 \equiv 1$ and $X_2 = X$.

Example: one-way layout We observe several samples from normal distributions (as in the "one-way ANOVA"). Model: $Y_{i,j} \sim N(\mu_j,\ \tau^{-1})$ for the i^{th} observation in sample j. Let us rename μ_j as β_j. Then we can write the model as

$$Y_{i,j} = \beta_j + \varepsilon_{i,j} \tag{1.5}$$

where $\varepsilon_{i,j} \sim N(0,\ \tau^{-1})$. Now, suppose that, instead of numbering the observations within each sample, we number them all in one long sequence $Y_1, \ldots, Y_n$, where $n = \sum_{j=1}^{J} n_j$. We need a way to indicate to which sample an observation belongs so we define regressors $X_1, \ldots, X_J$ where $x_{i,j} = 1$ if observation i is in sample j and $x_{i,j} = 0$ otherwise. Then our model is exactly of the form (1.4) if we set $p = J$.

Notice that, for fixed values of the regressors $x_{i,1}, \ldots, x_{i,p}$, (1.4) is linear in the coefficients $\beta_1, \ldots, \beta_p$. This is therefore called a *linear model* or a *linear regression*. It is called a *normal linear model* because of our assumption that the "errors" ε are normally distributed. The normal linear model includes a great variety of models which are commonly used in statistics. Generalisations and extensions allow an even greater variety but we will leave these for later. Just to illustrate that the linearity refers to the coefficients and not to the shape of a graph which we might draw to represent how Y changes, consider a model in which we want to describe the way that Y changes over time t using a cubic function of t. We simply write $x_{i,1} = 1,\ x_{i,2} = t_i,\ x_{i,3} = t_i^2$ and $x_{i,4} = t_i^3$. Then

$$Y_i = \beta_1 + \beta_2 t_i + \beta_3 t_i^2 + \beta_4 t_i^3 + \varepsilon_i.$$

Matrix notation

It is convenient to use matrix notation. We rewrite (1.4) as

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{1.6}$$

where $\underline{Y} = (Y_1, \ldots, Y_n)^T$ is a $n \times 1$ vector of observations on Y,

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

is a $n \times p$ matrix whose elements are known x-values. (In some cases all of the elements are 0 or 1). We call X the *design matrix*. This name reflects the fact that sometimes, that is is designed experiments, the elements of X are deliberately chosen and X then represents the *design* of the experiment. The $p \times 1$ vector of unknown parameters is $\underline{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\underline{\varepsilon} = (\varepsilon_1 \ldots, \varepsilon_n)^T$ is a $n \times 1$ error vector. The vector of random errors has a multivariate normal distribution (given τ):

$$\underline{\varepsilon} \sim N_n(\underline{0},\ \tau^{-1} I)$$

where $\underline{0}$ is a vector of zeroes and I is a $n \times n$ identity matrix.

Given τ and $\underline{\beta}$, the vector of observations $\underline{y}$ is an observation from a multivariate normal distribution:

$$\underline{Y} \sim_n N(X\underline{\beta},\ \tau^{-1} I).$$

Example: Regression on a single covariate Here $Y_i = \alpha + \beta x_i + \varepsilon_i$. We have

$$X = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_1 & x_2 & x_3 & \ldots & x_n \end{pmatrix}^T$$

and

$$\underline{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

Example: one-way layout (as above). Suppose, for illustration, that we have four samples, each with three observations. Then we have

$$\underline{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \qquad \text{and} \qquad X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

(There is, in fact, more than one way to *parameterise*, that is express in terms of parameters, this model and it is sometimes convenient to do it in a different way).

Notice that the design matrix contains one column corresponding to each of the coefficients $\beta_1, \ldots, \beta_p$.

1.2 Inference for the normal linear model

1.2.1 Likelihood and sufficient statistics

Given the model in (1.6) and a data vector $\underline{y}$ containing n observations, the likelihood is

$$L = (2\pi)^{-n/2}\tau^{n/2}\exp\left\{-\frac{\tau}{2}(\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta})\right\}.$$

We will assume in what follows that the design matrix X is of full rank and that therefore $(X^TX)^{-1}$ exists. If X is not of full rank then this does not mean that the likelihood does not exist nor that no Bayesian inference is possible. However it does mean that there is at least one linear function of $\underline{\beta}$ about which the data will tell us nothing. In such a case it may be best to reconsider the model. For example, suppose that, instead of the model in (1.5), we had $Y_{i,j} = \mu + \beta_j + \varepsilon_{i,j}$ where μ is meant to represent a sort of overall mean. Then, when we put this in the form (1.4), μ becomes, in effect, β_{J+1} and we have an extra regressor X_{J+1} where $x_{i,J+1} = 1$. However, for all i, $x_{i,J+1} = x_{i,1} + \cdots + x_{i,J}$ so the rank of X is still J, not $J+1$ even though it now has $J+1$ columns. It is easy to see that, in this case, we have too many parameters and they can not all be *identified*. If we replaced $\beta_1, \ldots, \beta_J$ with $\tilde{\beta}_1, \ldots, \tilde{\beta}_J$, where $\tilde{\beta}_j = \beta_j + \delta$, and β_{J+1} with $\tilde{\beta}_{J+1} = \beta_{J+1} - \delta$, then we would get exactly the same model and exactly the same likelihood so the data can not tell us about δ and therefore not about the complete set of values of $\beta_1, \ldots, \beta_J$. We could, however, learn about the differences $\beta_j - \beta_{J+1}$ for $j = 1, \ldots, J$.

So, assuming that $(X^TX)^{-1}$ exists, let us write

$$\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}.$$

We call $\hat{\underline{\beta}}$ the *least squares estimates* of $\underline{\beta}$.

Then

$$\begin{aligned}(\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta}) &= (\underline{y}-X\hat{\underline{\beta}}-X[\underline{\beta}-\hat{\underline{\beta}}])^T(\underline{y}-X\hat{\underline{\beta}}-X[\underline{\beta}-\hat{\underline{\beta}}]) \\ &= (\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}}) + (\underline{\beta}-\hat{\underline{\beta}})^TX^TX(\underline{\beta}-\hat{\underline{\beta}}) - 2(\underline{\beta}-\hat{\underline{\beta}})^TX^T(\underline{y}-X\hat{\underline{\beta}})\end{aligned}$$

but

$$(\underline{\beta}-\hat{\underline{\beta}})^TX^T(\underline{y}-X\hat{\underline{\beta}}) = (\underline{\beta}-\hat{\underline{\beta}})^T\{X^T\underline{y} - X^TX(X^TX)^{-1}X^T\underline{y}\} = 0.$$

Thus

$$L = (2\pi)^{-n/2}\tau^{n/2}\exp\left\{-\frac{\tau}{2}[S_d + (\underline{\beta}-\hat{\underline{\beta}})^TX^TX(\underline{\beta}-\hat{\underline{\beta}})]\right\} \tag{1.7}$$

and S_d and $\hat{\underline{\beta}}$ are sufficient for τ and $\underline{\beta}$, where

$$S_d = (\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}}).$$

Moreover, if τ is known, then $\hat{\underline{\beta}}$ is sufficient for $\underline{\beta}$.

The sampling distribution of $\underline{Y}$ is

$$\underline{Y} \mid \tau, \underline{\beta} \sim N_n(X\underline{\beta},\ \tau^{-1}I)$$

so the sampling distribution of $\hat{\underline{\beta}}$ is

$$\hat{\underline{\beta}} \mid \tau, \underline{\beta} \sim_p N(\underline{\beta},\ \tau^{-1}[X^TX]^{-1})$$

since $(X^TX)^{-1}X^TX\underline{\beta} = \underline{\beta}$ and $(X^TX)^{-1}X^T[\tau^{-1}I]X(X^TX)^{-1} = \tau^{-1}[X^TX]^{-1}$. Thus the "data precision" is τX^TX.

1.2.2 Inference with known error precision

Suppose that the error precision is known and that our prior distribution for $\underline{\beta}$ is a multivariate normal distribution with mean $\underline{b}_0$ and variance $V_0 = P_0^{-1}$. Then the posterior distribution is a multivariate normal distribution with mean $\underline{b}_1$ and variance $V_1 = P_1^{-1}$ where

$$\begin{aligned} \underline{b}_1 &= P_1^{-1}(P_0\underline{b}_0 + P_d\underline{\hat{\beta}}), \\ P_1 &= P_0 + P_d \\ \text{and} \qquad P_d &= \tau X^T X. \end{aligned}$$

The matrices P_0 and P_1 are the prior and posterior *precision matrices* respectively.

Proof: The prior density is proportional to

$$\exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{b}_0)^T P_0(\underline{\beta} - \underline{b}_0)\right\}.$$

The posterior density is therefore proportional to

$$\begin{aligned} h(\underline{\beta}) &= \exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{b}_0)^T P_0(\underline{\beta} - \underline{b}_0)\right\} \exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{\hat{\beta}})^T P_d(\underline{\beta} - \underline{\hat{\beta}})\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T(P_0 + P_d)\underline{\beta} - 2(\underline{b}_0^T P_0 + \underline{\hat{\beta}}^T P_d)\underline{\beta} + \underline{b}_0^T P_0\underline{b}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T(P_0 + P_d)\underline{\beta} - 2(\underline{b}_0^T P_0 + \underline{\hat{\beta}}^T P_d)(P_0 + P_d)^{-1}(P_0 + P_d)\underline{\beta} + \underline{b}_0^T P_0\underline{\beta}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T(P_0 + P_d)\underline{\beta} - 2\underline{b}_1^T(P_0 + P_d)\underline{\beta} + \underline{b}_0^T P_0\underline{b}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T(P_0 + P_d)\underline{\beta} - 2\underline{\beta}_1^T(P_0 + P_d)\underline{\beta} + \underline{b}_1^T(P_0 + P_d)\underline{b}_1\right]\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}\left[\underline{b}_0^T P_0\underline{b}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}} - \underline{b}_1^T(P_0 + P_d)\underline{b}_1\right]\right\} \\ &= \exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{b}_1)^T(P_0 + P_d)(\underline{\beta} - \underline{b}_1)\right\} \times \exp\left\{-\frac{1}{2}\left[\underline{b}_0^T P_0\underline{b}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}} - \underline{b}_1^T(P_0 + P_d)\underline{b}_1\right]\right\} \end{aligned}$$

which is proportional to

$$\exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{b}_1)^T(P_0 + P_d)(\underline{\beta} - \underline{b}_1)\right\}$$

which, in turn, is proportional to the pdf of a normal distribution with mean $\underline{b}_1$ and precision matrix $P_0 + P_d$.

Example 1

In a regression on a single covariate, where $Y_i = \alpha + \beta x_i + \varepsilon_i$, we have

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \qquad \text{and} \qquad X^T \underline{y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

So

$$(X^T X)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

where

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

and $\bar{x}$ is the sample mean of x, that is $(\sum_{i=1}^n x_i)/n$.
Therefore we can find the least squares estimates of $\underline{\beta} = (\alpha,\ \beta)^T$ as

$$\begin{aligned} \hat{\underline{\beta}} &= (X^T X)^{-1} X^T \underline{y} \\ &= \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{pmatrix} = \begin{pmatrix} \bar{y} - \bar{x} S_{xy}/S_{xx} \\ S_{xy}/S_{xx} \end{pmatrix}, \end{aligned}$$

where

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

and $\bar{y}$ is the sample mean of y, that is $(\sum_{i=1}^n y_i)/n$.

For the shoe-size and height example in section 1.1.2 we have

$$\underline{b}_0 = \begin{pmatrix} 55.7 \\ 1.7 \end{pmatrix} \qquad \text{and} \qquad P_0 = V_0^{-1} = \begin{pmatrix} 0.2222 & 1.7778 \\ 1.7778 & 16.2222 \end{pmatrix}.$$

Suppose that we choose $\tau = 0.5$, corresponding to an error standard deviation of 1.414 inches. From the data (a class of first-year students some years ago) we have $n = 152$, $\sum x_i = 1208.5$, $\sum y_i = 10427$, $\sum x_i^2 = 10202.25$ and $\sum x_i y_i = 83828$. From these we can calculate

$$\underline{\hat{\beta}} = \begin{pmatrix} 56.19516 \\ 1.56006 \end{pmatrix}$$

and

$$P_d = 0.5 \begin{pmatrix} 152 & 1208.5 \\ 1208.5 & 10202.25 \end{pmatrix} = \begin{pmatrix} 76.00 & 604.25 \\ 604.25 & 5101.125 \end{pmatrix}.$$

Therefore

$$\begin{aligned} P_1 &= P_0 + P_d = \begin{pmatrix} 76.2222 & 606.0278 \\ 606.0278 & 5117.3472 \end{pmatrix}, \\ V_1 &= P_1^{-1} = \begin{pmatrix} 0.22458 & -0.026597 \\ -0.026597 & 0.0033451 \end{pmatrix}, \\ \underline{b}_1 &= P_1^{-1}(P_0\underline{b}_0 + P_d\underline{\hat{\beta}}) = \begin{pmatrix} 56.1894 \\ 1.5610 \end{pmatrix}. \end{aligned}$$

The following table summarises the changes in our beliefs about α and β from prior to posterior.

	Prior		Posterior	
	Mean	Std. deviation	Mean	Std. deviation
α	55.7	6.042	56.189	0.474
β	1.7	0.707	1.561	0.058

The parameters are strongly negatively correlated in both the prior and posterior distributions. The prior correlation is -0.94 and the posterior correlation is -0.97.

Example 2

After certain material is extracted from an organism, the concentration of a certain compound in the material decreases exponentially over time.

Our model is

$$y_i = a - ct_i + \varepsilon_i$$

for $i = 1, \ldots, 6$, where y_i is the logarithm of a measurement on the concentration of the compound t_i minutes after the material is extracted. We assume that $\varepsilon_i \sim N(0, \sigma^2)$, where σ^2 is known to be 0.0025, and ε_i is independent of ε_j for $i \neq j$. So $\tau = 400$.
Our prior distribution for $\underline{\beta} = (a,\ c)^T$ is normal with mean

$$\underline{\beta}_0 = \begin{pmatrix} 4.605 \\ 0.01 \end{pmatrix}$$

and variance

$$V_0 = \begin{pmatrix} 0.1251 & 0 \\ 0 & 0.000025 \end{pmatrix}.$$

Hence

$$P_0 = \begin{pmatrix} 7.9936 & 0 \\ 0 & 40000 \end{pmatrix}.$$

The data are as follows.

i	1	2	3	4	5	6
Time t_i	25	50	75	100	125	150
Measured Concentration $\tilde{Z}$	113	81	74	52	43	36
Log Concentration Y	4.73	4.39	4.30	3.95	3.76	3.58

We can write $x_i = -t_i$. Then

$$X^T X = \begin{pmatrix} 6 & -525 \\ -525 & 56875 \end{pmatrix}.$$

Hence

$$P_1 = P_0 + \tau X^T X = \begin{pmatrix} 2407.994 & -210000 \\ -210000 & 22790000 \end{pmatrix}$$

and

$$V_1 = P_1^{-1} = \begin{pmatrix} 2.114458 \times 10^{-3} & 1.948382 \times 10^{-5} \\ 1.948382 \times 10^{-5} & 2.234139 \times 10^{-7} \end{pmatrix}.$$

The least squares estimates are

$$\underline{\hat{\beta}} = \begin{pmatrix} 4.91733 \\ 0.0091314 \end{pmatrix}.$$

The posterior mean of $(a, c)^T$ is

$$\underline{b}_1 = P_1^{-1}(P_0 \underline{b}_0 + P_d \underline{\hat{\beta}}) = \begin{pmatrix} 4.9127 \\ 0.0090905 \end{pmatrix}.$$

1.3 Inference with a conjugate prior

1.3.1 Prior and posterior

Suppose now that τ is unknown. There is a conjugate prior.

We give τ a gamma Ga$(d_0/2,\ d_0v_0/2)$ prior. Then $d_0v_0\tau$ has a $\chi^2_{d_0}$ distribution.

We then define the *conditional* prior distribution of $\underline{\beta}$ given τ as a multivariate normal distribution with mean $\underline{b}_0$ and precision $P_0 = C_0\tau$, where the value of C_0 is specified. Thus the prior precision of $\underline{\beta}$ is proportional to the error precision τ. It is easily shown that the marginal prior distribution of β_j is such that

$$\frac{\beta_j - b_{0,j}}{\sqrt{v_0/c_{0,j,j}}} \sim t_{d_0}$$

where $b_{0,j}$ is the j^{th} element of $\underline{b}_0$ and $c^{-1}_{0,j,j}$ is the j^{th} diagonal element of C_0^{-1}.

The prior density is then proportional to

$$\tau^{d_0/2-1}e^{-\tau(d_0v_0/2)}\tau^{p/2}\exp\left\{-\frac{\tau}{2}(\underline{\beta}-\underline{b}_0)^TC_0(\underline{\beta}-\underline{b}_0)\right\}.$$

From (1.7) the likelihood is proportional to

$$\tau^{n/2}\exp\left\{-\frac{\tau}{2}S_d\right\}\exp\left\{-\frac{\tau}{2}(\hat{\underline{\beta}}-\underline{\beta})^TC_d(\hat{\underline{\beta}}-\underline{\beta})\right\}$$

where $C_d = X^TX$.

The posterior density is therefore proportional to

$$\tau^{(d_0+n)/2-1}e^{-\tau(d_0v_0+S_d)/2}\tau^{p/2}\exp\left\{-\frac{\tau}{2}[(\underline{\beta}-\underline{b}_0)^TC_0(\underline{\beta}-\underline{b}_0)+(\hat{\underline{\beta}}-\underline{\beta})^TC_d(\hat{\underline{\beta}}-\underline{\beta})]\right\}.$$

Some further algebra shows that the posterior density is proportional to

$$\tau^{d_1/2-1}e^{-\tau d_1v_1/2}\tau^{p/2}|C_1|^{1/2}\exp\left\{-\frac{\tau}{2}(\underline{\beta}-\underline{b}_1)^TC_1(\underline{\beta}-\underline{b}_1)\right\}$$

where

$$\begin{aligned}
d_1 &= d_0+n\\
v_1 &= \frac{d_0v_0+nv_d}{d_0+n}\\
v_d &= \frac{S_d+R}{n}\\
R &= \underline{b}_0^TC_0\underline{b}_0+\hat{\underline{\beta}}^TC_d\hat{\underline{\beta}}-\underline{b}_1^TC_1\underline{b}_1\\
C_1 &= C_0+C_d\\
\underline{b}_1 &= (C_0+C_d)^{-1}(C_0\underline{b}_0+C_d\hat{\underline{\beta}})
\end{aligned}$$

Thus

- The marginal posterior distribution of τ is gamma Ga$(d_1/2,\ d_1v_1/2)$.
  So $d_1v_1\tau \sim \chi^2_{d_1}$.
- The conditional posterior distribution of $\underline{\beta}$ given τ is multivariate normal with mean $\underline{b}_1$ and variance $P_1^{-1} = \tau^{-1}C_1^{-1}$.

It is convenient to use a R function to do the calculations. A suitable function is shown in figure 1.1. The prior specification is supplied as a list containing $d_0,\ v_0,\ b_0$ and V_0, where $(V_0/v_0)^{-1} = C_0$. The function returns a list containing $d_1,\ v_1,\ b_1$ and V_1, where $(V_1/v_1)^{-1} = C_1$. The data are supplied as a matrix X and a vector $\underline{y}$. The function can be called with a command such as the following.

```
posterior<-linmod(prior,X,y)
```

```
linmod<-function(prior,X,y,unknown=TRUE)
{Xt<-t(X)
 Cd<-Xt%*%X
 Xty<-Xt%*%y
 b0<-prior$b
 betahat<-solve(Cd,Xty)
 n<-length(y)
 C0<-solve(prior$V/prior$v)
 C1<-C0+Cd
 b1<-solve(C1,(C0%*%b0+Cd%*%betahat))
 res<-y-X%*%betahat
 Sd<-sum(res^2)
 if (unknown)
    {d1<-prior$d+n
     R<-t(b0)%*%C0%*%b0 + t(betahat)%*%Cd%*%betahat - t(b1)%*%C1%*%b1
     nvd<-Sd+R
     v1<-(prior$d*prior$v + nvd)/d1
     v1<-v1[1,1]
     }
 else
    {v1<-prior$v
     d1<-0
     }
 V1<-v1*solve(C1)
 result<-list(d=d1,v=v1,b=b1,V=V1)
 result
 }
```

Figure 1.1: R function for the normal linear model

Optionally, we can use a command such as the following.

```
posterior<-linmod(prior,X,y,unknown=FALSE)
```

In this latter case the calculations for the known-τ case are used and the `prior` argument is a list containing $v_0 = \tau^{-1}$, b_0 and $V_0 = P_0^{-1}$. Similarly the result is a list containing $v_1 = v_0 = \tau^{-1}$, b_0 and $V_1 = P_1^{-1}$. The result in this case also contains the value $d_1 = 0$.

The use of the function is illustrated in the following examples.

Example 1

This is the example involving shoe sizes and heights of students, as in section 1.2.2. The only difference here is that we make τ unknown with $d_0 = 2$ and $v_0 = 2$. We can use the R function as follows.

```
> Xshoe<-matrix(c(rep(1,152),shoesize),ncol=2)
> d0shoe<-2
> v0shoe<-2
> b0shoe<-matrix(c(55.7,1.7),ncol=1)
> V0shoe<-matrix(c(36.5,-4,-4,0.5),ncol=2)
> priorshoe<-list(d=d0shoe,v=v0shoe,b=b0shoe,V=V0shoe)
> postshoe<-linmod(priorshoe,Xshoe,height)
> postshoe
$d
[1] 154

$v
```

Treatment	Diet	Weight gain									
1	Beef Low	90	76	90	64	86	51	72	90	95	78
2	Beef High	73	102	118	104	81	107	100	87	117	111
3	Cereal Low	107	95	97	80	98	74	74	67	89	58
4	Cereal High	98	74	56	111	95	88	82	77	86	92

Table 1.1: Weight gains in rats given different diets

```
[1] 3.515502

$b
          [,1]
[1,] 56.189352
[2,]  1.561022

$V
           [,1]          [,2]
[1,]  0.3947624 -0.046750199
[2,] -0.0467502  0.005879935
```

The posterior means are exactly the same as in the known-τ case. This is a property of the conjugate prior when it is specified in this way, with everything unchanged and v_0 equal to the previous "known" value. It seems though that the error variance may be a little greater than our "known" value.

Example 2

The data in table 1.1 are from Snedecor and Cochran (1967) and are also given by Hand *et al.* (1994). They give the gains in weight of rats fed on four different diets. The diets differ in terms of the amount of protein ("low" or "high") and the source of the protein ("beef" or "cereal").

Suppose that our prior beliefs are as follows. Given parameters $\underline{\mu} = (\mu_1, \ldots, \mu_4)^T$, τ, the weight gains $Y_{1,1}, \ldots, Y_{10,4}$ are independent with $Y_{i,j} \sim N(\mu_j, \tau^{-1})$. Our prior distribution for τ is gamma $\mathrm{Ga}(d_0/2,\ d_0 v_0/2)$ with $d_0 = 2$ and $v_0 = 60$. Our conditional prior distribution for $\underline{\mu}$ is $N_4(\underline{M}_0, (\tau C_0)^{-1})$ with $\underline{M}_0 = (80, 80, 80, 80)^T$ and

$$C_0 = \frac{1}{8}\begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}^{-1} = \frac{1}{40}\begin{pmatrix} 4 & -1 & -1 & -1 \\ -1 & 4 & -1 & -1 \\ -1 & -1 & 4 & -1 \\ -1 & -1 & -1 & 4 \end{pmatrix}.$$

Consider an alternative way of formulating this example. Instead of working directly in terms of the four means $\mu_1, \ldots, \mu_4$, we can use different parameters. We can write

$$\begin{aligned} \mu_1 &= \mu - \beta_a - \beta_s + \gamma, \\ \mu_2 &= \mu + \beta_a - \beta_s - \gamma, \\ \mu_3 &= \mu - \beta_a + \beta_s - \gamma, \\ \mu_4 &= \mu + \beta_a + \beta_s + \gamma. \end{aligned}$$

Here μ is an *overall mean*, β_a is an *effect* due to the amount of protein, β_s is an effect due to the source of protein. The interaction effect allows the treatment means to be unrestricted. It allows for the mean for, eg., "cereal high" not to be obtained simply by adding the source effect and the amount effect to the overall mean. It is easily seen that

$$\underline{\beta} = (\mu,\ \beta_a,\ \beta_s,\ \gamma)^T = H\underline{\mu}$$

where

$$H = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

So, if our prior mean and conditional prior variance for $\underline{\mu}$ were $\underline{M}_0$ and $(\tau C_{0,\mu})^{-1}$ respectively, then our prior mean and prior variance for $\underline{\beta}$ are

$$\underline{b}_0 = H\underline{M}_0 = \begin{pmatrix} 80 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad (\tau C_0)^{-1} = H(\tau C_{0,\mu})^{-1}H^T = \tau^{-1}\begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

Hence

$$C_0 = \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{pmatrix}.$$

In practice we might assess the prior distribution for $\underline{\beta}$ directly rather than through a prior distribution for $\underline{\mu}$. Also we might well wish to give γ a smaller prior variance than β_a or β_s since we might judge that such an interaction effect is likely to be less important than the *main* effects of amount and source of protein.

Here is the calculation of the posterior distribution using the R function `linmod`. The vector `ratgain` contains the weight gains $(90, 76, 90, \ldots, 86, 92)$.

```
> x1<-rep(1,40)
> x2<-rep(c(-1,1,-1,1),c(10,10,10,10))
> x3<-rep(c(-1,-1,1,1),c(10,10,10,10))
> x4<-rep(c(1,-1,-1,1),c(10,10,10,10))
> Xrat<-cbind(x1,x2,x3,x4)
> d0rat<-2
> v0rat<-60
> b0rat<-matrix(c(80,0,0,0),ncol=1)
> V0rat<-60*diag(c(10,2,2,2))
> priorrat<-list(d=d0rat,v=v0rat,b=b0rat,V=V0rat)
> postrat<-linmod(priorrat,Xrat,ratgain)
> postrat
$d
[1] 42

$v
[1] 195.3410

$b
        [,1]
x1 87.231920
x2  5.629630
x3 -2.320988
x4 -4.641975

$V
         x1       x2       x3       x4
x1 4.871347 0.000000 0.000000 0.000000
x2 0.000000 4.823235 0.000000 0.000000
x3 0.000000 0.000000 4.823235 0.000000
x4 0.000000 0.000000 0.000000 4.823235
```

To find a posterior credible interval for the j^{th} element of $\underline{\beta}$ we can use $b_{1,j} \pm t\sqrt{v_1/c_{1,j,j}}$ where $b_{1,j}$ is the j^{th} element of $\underline{b}_1$, t is an appropriate quantile of the Student's t-distribution on d_1 degrees of freedom and $c^{-1}_{1,j,j}$ is the j^{th} diagonal element of C_1^{-1}. In this example V_1 and C_1 are diagonal so the j^{th} diagonal element of V_1 is, in fact, $v_1/c_{1,j,j}$. The 97.5% point of the t_{42} distribution is 2.018 so, for example, a 95% interval for μ is $87.23192 \pm 2.018 \times \sqrt{4.871347}$. The following table gives the posterior means and 95% posterior intervals for the elements of $\underline{\beta}$.

Parameter		Mean	95% Interval	
μ	Overall mean	87.2319	82.7776	91.6862
β_a	Amount effect	5.6296	1.1975	10.0617
β_s	Source effect	-2.3210	-6.7531	2.1111
γ	Interaction effect	-4.6420	-9.0741	-0.2099

It looks as though the most important effect may be amount of protein but, because of the interaction effect, the difference in weight gain between "low" and "high" amounts may be less when the source is cereal than when it is beef.

1.3.2 Linear functions of coefficients

In our posterior distribution $\tau \sim \text{Ga}(d_1/2,\ d_1 v_1/2)$ and $\beta \mid \tau \sim N_p(\underline{b}_1,\ (\tau C_1)^{-1})$.

Suppose that we are interested in some linear function of $\underline{\beta}$. For example, with $\underline{\beta} = (\beta_1, \beta_2, \beta_3)^T$, we might be interested in $\delta = \underline{x}\underline{\beta} = 4\beta_1 + 3\beta_2 - 5\beta_3$. This is, of course, the mean of Y when $\underline{x} = (4, 3, -5)$.

Then

$$\delta \mid \tau \sim N(\underline{x}\underline{b}_1,\ \underline{x}(\tau C_1)^{-1}\underline{x}^T).$$

That is

$$\delta \mid \tau \sim N(\underline{x}\underline{b}_1,\ (\tau c_{\delta,1})^{-1})$$

where $c^{-1}_{\delta,1} = \underline{x}C_1^{-1}\underline{x}^T$.

So the marginal posterior for δ is such that

$$\frac{\delta - \underline{x}\underline{b}_1}{\sqrt{v_1/c_{\delta,1}}} = \frac{\delta - \underline{x}\underline{b}_1}{\sqrt{\underline{x}V_1\underline{x}^T}} \sim t_{d_1},$$

where $V_1 = v_1 C_1^{-1}$.

1.3.3 Prediction

Very often our purpose in using a regression is to be able to make predictions. That is, we want to find the distribution of a future observation on the dependent variables, or perhaps a collection of future observations. In the case of the normal linear model this is usually straightforward.

Suppose that we are going to make a new observation on Y and the covariate values will be $x_{0,1}, \ldots, x_{0,p}$. We arrange these covariate values into a vector $\underline{x}_0$. For convenience, we regard this as a *row* vector rather than the more usual column vector. That is, its dimension is $(1 \times p)$ rather than $(p \times 1)$. Then we can write

$$Y = \underline{x}_0 \underline{\beta} + \varepsilon,$$

where ε is a *new* error which is conditionally independent of any data which we have observed, given τ, and therefore also conditionally independent of the unknown value of $\underline{\beta}$, given τ. Given τ, the distribution of ε is $N(0,\ \tau^{-1})$. Let us assume that our distribution for $\underline{\beta}$ is normal. Then, given τ, the distribution of Y is normal with mean given by the mean of $\underline{x}_0 \underline{\beta}$ and variance given by the sum of the variance of $\underline{x}_0 \underline{\beta}$ and the variance of ε.

Suppose that we are making a *posterior* prediction. That is, we are making our prediction after we have observed some data and our conditional posterior distribution for $\underline{\beta} \mid \tau$ is $N(\underline{b}_1,\ \tau^{-1} C_1^{-1})$. Then we can write

$$\begin{aligned} Y \mid \tau &\sim N(\underline{x}_0 \underline{b}_1,\ \underline{x}_0 [C_1 \tau]^{-1} \underline{x}_0^T + \tau^{-1}) \\ &\sim N(\underline{x}_0 \underline{b}_1,\ [c_p \tau]^{-1}), \end{aligned}$$

where

$$c_p = \{1 + \underline{x}_0 C_1^{-1} \underline{x}_0^T\}^{-1}.$$

In the conjugate case, where our posterior distribution for τ is $\tau \sim \mathrm{Ga}(d_1/2,\ d_1 v_1/2)$, it follows that the marginal distribution for Y is given by

$$\frac{Y - \underline{x}_0 \underline{b}_1}{\sqrt{v_1 / c_p}} = \frac{Y - \underline{x}_0 \underline{b}_1}{\sqrt{v_1 + \underline{x}_0 V_1 \underline{x}_0^T}} \sim t_{d_1}. \tag{1.8}$$

This is our predictive distribution for the new observation Y. It includes both the uncertainty due to our lack of knowledge of the model parameters and our uncertainty associated with the new error.

More generally we might want a joint predictive distribution for a vector $\underline{Y}$ of new observations with sampling distribution $N_m(X_0 \beta,\ \tau^{-1} I)$, where I is an identity matrix and the covariate values for the i^{th} element of $\underline{Y}$ give the i^{th} row of X_0. Then

$$\begin{aligned} \underline{Y} \mid \tau &\sim N_m(X_0 \underline{b}_1,\ X_0 [C_1 \tau]^{-1} X_0^T + \tau^{-1} I) \\ &\sim N_m(X_0 \underline{b}_1,\ [C_p \tau]^{-1}), \end{aligned}$$

where

$$C_p = \{I + X_0 C_1^{-1} X_0^T\}^{-1}.$$

Example

Consider Example 1 of section 1.3.1. Suppose that we want to predict the height of a student with shoe size 10. Then $\underline{x}_0 = (1, 10)$ and the mean of our predictive distribution is $\underline{x}_0\underline{b}_1 = 56.189352 + 10 \times 1.561022 = 71.799572$. Now $C_1^{-1} = V_1/v_1$ so $c_p = \{1 + \underline{x}_0 C_1^{-1}\underline{x}_0^T\}^{-1} = v_1\{v_1 + \underline{x}_0 V_1 \underline{x}_0^T\}^{-1}$. Hence

$$\frac{v_1}{c_p} = v_1 + \underline{x}_0 V_1 \underline{x}_0^T = 3.515502 + 0.04775197 = 3.563254$$

and

$$\frac{Y - 71.799572}{\sqrt{3.563254}} \sim t_{154}.$$

The upper 95% point of the t_{154} distribution is 1.654808 so a 90% predictive interval for Y is given by $71.799572 \pm 1.654808 \times \sqrt{3.563254}$. That is $68.7 < Y < 74.9$.

1.3.4 Other cases

We have looked in detail at the conjugate case. We can also analyse linear models with a semi-conjugate prior or with a non-conjugate prior. In the semi-conjugate case we need numerical integration in one dimension, that of τ. In the non-conjugate case we usually need more difficult numerical integration and it is usually easier to use MCMC.

1.4 Practical 1

1.4.1 Abrasion Loss

The data in Table 1.2 are taken from Davies and Goldsmith (1972). They come from an experiment to investigate how the resistance of rubber to abrasion is affected by other properties. These are X_1, its hardness, in degrees Shore, and X_2, its tensile strength (in kg per square cm). The dependent variable Y is abrasion loss in g per hour. This is the weight loss due to abrasion which was measured over a fixed time.

You are to fit a linear regression of Y on X_1 and X_2. The model is

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

where, given τ, the errors ε_i are independent with $\varepsilon_i \sim N(0, \tau^{-1})$.

1. Install the function `linmod`. It is available from the Web page at

Abrasion loss Y	Hardness X_1	Tensile strength X_2
372	45	162
206	55	233
175	61	232
154	66	231
136	71	231
112	71	237
55	81	224
45	86	219
221	53	203
166	60	189
164	64	210
113	68	210
82	79	196
32	81	180
228	56	200
196	68	173
128	75	188
97	83	161
64	88	119
249	59	161
219	71	151
186	80	165
155	82	151
114	89	128
341	51	161
340	59	146
283	65	148
267	74	144
215	81	134
148	86	127

Table 1.2: Abrasion loss data

```
http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/
```

You can install it by copying and pasting or by

```
source("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/linmod.txt")
```

2. The data are available in the file `abrasion.txt` which is available from the Web page. You can read the data using commands such as the following.

```
abrasion<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/abrasion.txt")
loss<-abrasion[,1]
hard<-abrasion[,2]
tens<-abrasion[,3]
```

3. Construct a design matrix X as follows.

```
X<-matrix(c(rep(1,30),hard,tens),ncol=3)
```

4. Our prior distribution is as follows. We give τ a $\mathrm{Ga}(d_0/2,\ d_0v_0/2)$ distribution with $d_0 = 4$ and $v_0 = 1600$. Conditional on τ we give $\underline{\beta} = (\beta_0, \beta_1, \beta_2)^T$ a multivariate normal prior distribution with mean vector $\underline{b}_0 = (150, 0, 0)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and we construct V_0 as follows. Consider first a reference value with $x_1 = 60$ and $x_2 = 200$. If we consider the model $\mathrm{E}(Y) = \tilde{\beta}_0 + \beta_1(x_1 - 60) + \beta_2(x_2 - 200)$ we obtain for the parameters $\underline{\tilde{\beta}} = (\tilde{\beta}_0,\ \beta_1,\ \beta_2)^T$ the matrix

$$\tilde{V}_0 = 1600 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.25 \end{pmatrix}.$$

Now, since $\underline{\beta} = H\underline{\tilde{\beta}}$ where

$$H = \begin{pmatrix} 1 & -60 & -200 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

we can construct $V_0 = H\tilde{V}_0H^T$ as follows.

```
V0tilde<-matrix(c(1600,0,0,0,400,0,0,0,400),ncol=3)
H<-matrix(c(1,0,0,-60,1,0,-200,0,1),ncol=3)
V0<-H%*%V0tilde%*%t(H)
```

Put all of the elements of the prior together.

```
d0<-4
v0<-1600
b0<-matrix(c(150,0,0),ncol=1)
priorabloss<-list(d=d0,v=v0,b=b0,V=V0)
```

5. Find the posterior.

```
postabloss<-linmod(priorabloss,X,loss)
```

6. Find a 95% posterior predictive interval for the abrasion loss in a new observation with $x_1 = 80$ and $x_2 = 150$.

Ayrshire		Canadian	
Mature	2-yr-old	Mature	2-yr-old
3.74	4.44	3.92	4.29
4.01	4.37	4.95	5.24
3.77	4.25	4.47	4.43
3.78	3.71	4.28	4.00
4.10	4.08	4.07	4.62
4.06	3.90	4.10	4.29
4.27	4.41	4.38	4.85
3.94	4.11	3.98	4.66
4.11	4.37	4.46	4.40
4.25	3.53	5.05	4.33

Table 1.3: Butterfat percentages in milk

```
v1<-postabloss$v
V1<-postabloss$V
x0<-matrix(c(1,80,150),nrow=1)
mean<-x0%*%postabloss$b
var<-v1+x0%*%V1%*%t(x0)
tval<-qt(0.975,postabloss$d)
mean-tval*sqrt(var)
mean+tval*sqrt(var)
```

1.4.2 Butterfat

Table 1.3 shows part of a set of data taken from Sokal and Rohlf (1981). The table shows average butterfat percentages in the milk of forty cows. Twenty of the cows belong to each of two breeds, Ayrshire and Canadian. Within each breed, ten of the cows are mature (i.e. at least five years old) and ten are two-year-olds.

We adopt the following model.

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i$$

where Y_i is the butterfat percentage for cow i. We make the usual assumptions about $\varepsilon_1, \ldots, x_{40}$. That is, given τ, they are independent and $\varepsilon_i \sim N(0, \tau^{-1})$. The explanatory variables are as follows.

- Breed, where $x_{i,1} = -1$ if the breed of cow i is Ayrshire and $x_{i,1} = 1$ if the breed of cow i is Canadian.
- Age, where $x_{i,2} = -1$ if cow i is mature and $x_{i,2} = 1$ if cow i is a 2-year-old.
- Breed by age interaction, $x_{i,3} = x_{i,1}x_{i,2}$.

1. If you have not already done so, install the function `linmod` . (See above).

   ```
   source("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/linmod.txt")
   ```

2. The data are available in the file `butter.txt` which is available from the Web page.

   You can read the data using commands such as the following.

   ```
   butter<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/butter.txt")
   butter<-c(butter[,1],butter[,2],butter[,3],butter[,4])
   ```

3. Construct a design matrix X as follows.

```
z<-rep(10,4)
x0<-rep(1,40)
x1<-rep(c(-1,-1,1,1),z)
x2<-rep(c(-1,1,-1,1),z)
x3<-x1*x2
X<-cbind(x0,x1,x2,x3)
```

4. Our prior distribution is as follows. We give τ a $\mathrm{Ga}(d_0/2,\ d_0v_0/2)$ distribution with $d_0 = 6$ and $v_0 = 0.1$. Conditional on τ we give $\underline{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ a multivariate normal prior distribution with mean vector $\underline{b}_0 = (4, 0, 0, 0)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

$$V_0 = \begin{pmatrix} 40 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 2.5 \end{pmatrix}.$$

   We can construct V_0 as follows.

```
V0<-diag(c(40,10,10,2.5))
```

   Put all of the elements of the prior together.

```
d0<-6
v0<-0.1
b0<-matrix(c(4,0,0,0),ncol=1)
priorbutter<-list(d=d0,v=v0,b=b0,V=V0)
```

5. Find the posterior.

```
postbutter<-linmod(priorbutter,X,butter)
```

6. Find a 90% posterior interval for the mean butterfat percentage for 2-yr-old Ayrshire cows.

```
v1<-postbutter$v
V1<-postbutter$V
x0<-matrix(c(1,-1,1,-1),nrow=1)
mean<-x0%*%postbutter$b
var<-x0%*%V1%*%t(x0)
tval<-qt(0.95,postbutter$d)
mean-tval*sqrt(var)
mean+tval*sqrt(var)
```

1.5 Exercises

1. Table 1.4 shows the heights and weights of thirty eleven-year-old girls attending Heaton Middle School, Bradford. The data are taken from Open University (1983).

   The data are available in the file `height.txt` on the Web page.

   You can read the data using commands such as the following.

```
eleven<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/height.txt")
height<-eleven[,1]
weight<-eleven[,2]
```

Height (cm)	Weight (kg)	Height (cm)	Weight (kg)
135	26	133	31
146	33	149	34
153	55	141	32
154	50	164	47
139	32	146	37
131	25	149	46
149	44	147	36
137	31	152	47
143	36	140	33
146	35	143	42
141	28	148	32
136	28	149	32
154	36	141	29
151	48	137	34
155	36	135	30

Table 1.4: Heights and weights of eleven-year-old girls

(a) You should work in terms of the logarithms of both height and weight. So, let Y be the natural logarithm of the weight and X be the natural logarithm of the height. Calculate these and plot a graph to show the data.

Our model is

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

where, given the value of τ, the errors ε_i are independent and $\varepsilon_i \sim N(0, \tau^{-1})$.

(b) Our prior distribution is as follows. We give τ a $\text{Ga}(d_0/2,\ d_0 v_0/2)$ distribution with $d_0 = 6$ and $v_0 = 0.02$. Conditional on τ we give $\underline{\beta} = (\alpha, \beta)^T$ a bivariate normal prior distribution with mean vector $\underline{b}_0 = (-10, 3)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

$$V_0 = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}.$$

Find the posterior distribution. (I.e. explain it as I have explained the prior distribution but with the appropriate parameter values).

(c) Find a 95% posterior predictive interval for the natural logarithm of the weight of an eleven-year-old girl who is 145 cm tall and, convert this into a 95% posterior predictive interval for the actual weight of such a girl.

2. Table 1.5 gives some data from Till (1974). They give measured salinity values (parts per thousand) for three separate water masses in the Bimini Lagoon in the Bahamas.

The data are available in the file `salinity.txt` on the Web page.

You can read the data using commands such as the following.

```
bimini<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/salinity.txt")
salinity<-bimini[,1]
location<-bimini[,2]
mass1<-ifelse((location==1),1,0)
mass2<-ifelse((location==2),1,0)
mass3<-ifelse((location==3),1,0)
```

(a) Our model is

I	II	III
37.54	40.17	39.04
37.01	40.80	39.21
36.71	39.76	39.05
37.03	39.70	38.24
37.32	40.79	38.53
37.01	40.44	38.71
37.03	39.79	38.89
37.70	39.38	38.66
37.36		38.51
36.75		40.08
37.45		
38.85		

Table 1.5: Salinity measurements (parts per thousand)

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i$$

where, given the value of τ, the errors ε_i are independent and $\varepsilon_i \sim N(0, \tau^{-1})$ and $x_{i,j} = 1$ if observation i is from location j with $x_{i,j} = 0$ otherwise.

(b) Our prior distribution is as follows. We give τ a Ga$(d_0/2,\ d_0 v_0/2)$ distribution with $d_0 = 4$ and $v_0 = 0.3$. Conditional on τ we give $\underline{\beta} = (\beta_1, \beta_2, \beta_3)^T$ a multivariate normal prior distribution with mean vector $\underline{b}_0 = (40, 40, 40)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

$$V_0 = H \tilde{V}_0 H^T$$

where

$$\tilde{V}_0 = \begin{pmatrix} 40 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 25 \end{pmatrix}$$

and

$$H = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Find the posterior distribution. (I.e. explain it as I have explained the prior distribution but with the appropriate parameter values).

(c) Find a 95% posterior interval for the difference in mean salinity between water mass I and water mass II.

Note that, as an alternative to using the function `linmod` in this question, you could use the function `oneway` which is also available from the Web page.

1.6 Problems 3

Solutions to all questions are to be submitted in the Homework Letterbox no later than 4.00pm on Wednesday November 28th. Please note that you should give some attention to the presentation of your work. Describe the data, model, prior etc. and explain what you have done. Comment on your conclusions. A listing of the output from a R session with one or two things written on it will not get a very good mark on its own.

In questions 2 and 3, each student is given different data. For this purpose each student is given a reference number according to the table below. Please use the correct data and write your reference number on your work. In these questions you may, of course, use R functions such as `linmod` *for calculations.*

Reference numbers

Problems

1. Prior Elicitation

   Some household contents insurance policies require an estimate to be made of what it would cost to replace the existing contents. Suppose that a person has a large collection of books. We might attempt to predict the replacement cost of all of the books by looking at a sample. We might improve this prediction by taking into account an auxiliary variable such as the width of the spine of the book. (We might also distinguish between hardback and paperback books so suppose that we are only considering hardback books). Let C_i be the replacement cost, in £, of book i, and let w_i be its spine width in mm.

   Let

   $$Y_i = \log_e(C_i) \qquad \text{and} \qquad x_i = \log_e(w_i).$$

   It is believed that Y is related to X by

   $$Y_i = \alpha + \beta x_i + \varepsilon_i$$

   where Y_i and x_i refer to book i for $i = 1, \ldots, n$, $\varepsilon_i \sim N(0, \tau^{-1})$ and $\varepsilon_1, \ldots, \varepsilon_n$ are conditionally independent given τ.

   We give α and β a bivariate normal prior distribution. Find the parameters of this distribution based on the following prior judgments.

   Suppose that we could observe a very large number of books, each of which has a spine $w = 20$mm wide, and a very large number of books, each of which has a spine $w = 30$mm wide. Let the median replacement costs at these two spine widths be M_{20} and M_{30} respectively. Our prior median for M_{20} is 25 and our prior median for M_{30} is 35. Our prior upper quartile for M_{20} is 40 and our prior upper quartile for M_{30} is 55.

   Let

   $$m_{20} = \log_e(M_{20}) \quad \text{and} \quad m_{30} = \log_e(M_{30}).$$

   Our prior correlation for m_{20} and m_{30} is 0.75.

   Find the prior means, prior variances and prior covariance of $\alpha,\ \beta$.

   (10 marks)

2. Lowering blood pressure during surgery

It is sometimes necessary to lower a patient's blood pressure during surgery, using a hypotensive drug. The length of time over which the drug is administered varies and therefore so does the total dose. This, in turn, might affect the time it takes for the patient's blood pressure to return to normal.

The data provided are as follows, for $n = 53$ patients.

- The natural logarithm of the recovery time, T, in minutes.
- The natural logarithm of the dose, d, in milligrams.
- The average systolic blood pressure, b, in millimetres of mercury, during administration.

Let $Y = \ln(T)$, $x_1 = \ln(d) - 5$ and $x_2 = b - 60$. We will use a regression model with

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where $\beta_0, \beta_1, \beta_2$ are unknown parameters and, conditional on the values of the parameters, $\varepsilon_1, \ldots, \varepsilon_{53}$ are independent with $\varepsilon_i \sim N(0, \tau^{-1})$.

Our prior distribution is as follows.

We give τ a gamma prior, $\tau \sim \text{Ga}(1.5,\ 0.6)$. Conditional on τ we give $\underline{\beta} = (\beta_0, \beta_1, \beta_2)^T$ a multivariate normal prior distribution with mean vector $\underline{b}_0 = (3.0, -0.03, 0.5)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

$$V_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 10^{-4} & 0 \\ 0 & 0 & 0.04 \end{pmatrix}.$$

You can read the data using a command such as the following.

```
surgery<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/surgerydata.txt")
```

There are thirty columns.

- The log doses $\ln(d)$ are in column 1.
- The blood pressures b are in column 2.
- Your log recovery times t are in the column corresponding to your reference number. For example, if your reference number is 20 then your data are in column 20.

(a) Find the posterior distribution of $\beta_0, \beta_1, \beta_2, \tau$ (in the same form as the prior distribution).

(4 marks)

(b) Find and plot the posterior predictive probability density of the logarithm of the recovery time for a patient with log dose 4.0 and blood pressure 70 during administration.

(4 marks)

(c) Find and plot the posterior predictive probability density of the recovery time for a patient with log dose 4.0 and blood pressure 70 during administration.

(4 marks)

(d) Present, explain and comment on your findings clearly.

(8 marks)

Hint: You can use the following R commands to build the design matrix.

```
x1<-surgery[,1]-5
x2<-surgery[,2]-60
x0<-rep(1,53)
X<-cbind(x0,x1,x2)
```

3. Yields of barley

An experiment was conducted to investigate the effect of manure on the yield of barley. Four different levels of manure were compared: 1: no manure, 2: 0.01 tons per acre, 3: 0.02 tons per acre, 4: 0.04 tons per acre. Three different varieties of barley were used. The experimental plots were arranged in six blocks. (A "block" is an area of land).

You can read the data using a command such as the following.

```
barley<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/splitdata.txt")
```

There are thirty columns. Your barley yields y are in the column corresponding to your reference number. For example, if your reference number is 20 then your data are in column 20. Columns 1, 2 and 3 contain the block, variety and manure level respectively.

You can construct a suitable design matrix using the following R commands.

```
block<-barley[,1]
variety<-barley[,2]
manure<-barley[,3]
X<-matrix(nrow=72,ncol=11)
for (col in 1:4)
    {X[,col]<-ifelse(manure==col,1,0)
     }
b<-rep(12,6)
X[,5]<-rep(c(1, 1, 1,-1,-1,-1),b)
X[,6]<-rep(c(2,-1,-1, 0, 0, 0),b)
X[,7]<-rep(c(0, 1,-1, 0, 0, 0),b)
X[,8]<-rep(c(0, 0, 0, 2,-1,-1),b)
X[,9]<-rep(c(0, 0, 0, 0, 1,-1),b)
v<-rep(4,3)
x<-rep(c(2,-1,-1),v)
X[,10]<-rep(x,6)
x<-rep(c(0, 1,-1),v)
X[,11]<-rep(x,6)
```

The first four columns of X correspond to the four levels of manure. Columns 5-9 are for the block effects. (There are five degrees of freedom between the six blocks). Columns 10-11 are for the variety effects. (There are two degrees of freedom between the three varieties). (We could also fit interaction effects but we will leave that for now).

Let the parameters corresponding to the eleven columns of X be $\beta_1, \ldots, \beta_{11}$. Then the mean yield, $\mu_{m,b,v}$, for manure level m in block b with variety v is defined as follows.

$$\mu_{m,b,v} = \beta_m + \sum_{j=5}^{9} \beta_j w_{b,j} + \sum_{j=10}^{11} \beta_j z_{v,j}$$

Here $w_{b,j}$ and $z_{v,j}$ are defined as follows.

$w_{b,j}$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$
$b=1$	1	2	0	0	0
$b=2$	1	−1	1	0	0
$b=3$	1	−1	−1	0	0
$b=4$	−1	0	0	2	0
$b=5$	−1	0	0	−1	1
$b=6$	−1	0	0	−1	−1

$z_{v,j}$	$j=10$	$j=11$
$v=1$	2	0
$v=2$	−1	1
$v=3$	−1	−1

The actual yield for for manure level m in block b with variety v is

$$y_{m,b,v} = \mu_{m,b,v} + \varepsilon_{m,b,v}$$

where $\varepsilon_{m,b,v} \sim N(0, \tau^{-1})$ and $\varepsilon_{m,b,v}$ is independent of $\varepsilon_{m',b',v'}$ unless $(m,b,v) = (m',b',v')$.

Our prior distribution is as follows.

We give τ a gamma prior, $\tau \sim \text{Ga}(d_0/2,\ d_0v_0/2)$ with $d_0 = 2.1$ and $v_0 = 250$. Conditional on τ we give $\underline{\beta} = (\beta_0, \ldots, \beta_{11})^T$ a multivariate normal prior distribution with mean vector

$$\underline{b}_0 = 100(1,1,1,1,0,0,0,0,0,0,0,0)^T$$

and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

$$V_0 = \frac{1}{6}\begin{pmatrix} 24 & 12 & 12 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 24 & 12 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 12 & 24 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 12 & 12 & 24 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}.$$

You can construct V_0 in R, for example using the following commands.

```
V0<-matrix(0,nrow=11,ncol=11)
V0[1:4,1:4]<-matrix(2,nrow=4,ncol=4)+diag(2,4)
V0[5:11,5:11]<-diag(c(2,2,6,2,6,1,3))/6
```

(a) Find the posterior distribution of $\beta_0, \ldots, \beta_{11}, \tau$ (in the same form as the prior distribution).

(6 marks)

(b) Find a symmetric 95% posterior interval for the mean yield for Manure level 1 in Block 1 with Variety 1.

(6 marks)

(c) Present, explain and comment on your findings clearly.

(8 marks)

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	$\mathrm{Bin}(n, p)$	$\mathrm{var}(Y_i) = \mu(1 - \mu/n) \quad (\mu = np)$
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?

Chapter 3

Missing Data and Data Augmentation

3.1 Introduction to Graphical Models

In the rest of the module it will sometimes be useful to use graphical representations of models. We will look at a particular type of graph called a *directed acyclic graph* or *dag*.

Suppose that we are going to observe a number of animals of the same species. Each animal might or might not have a particular gene. Suppose that all animals in the population are considered to be exchangeable with respect to having this gene. Let $T_i = 1$ if animal i has the gene. Otherwise $T_i = 0$. Because of the exchangeability, we can represent the relationships in our beliefs about $T_1, T_2, \ldots$ by introducing θ to represent the unknown overall proportion of animals in the population which have the gene. The graph for two animals is shown in Figure 3.1.

With three unknowns, A, B, C, we can always write the joint probability as

$$\Pr(A, B, C) = \Pr(A)\Pr(B|A)\Pr(C|A, B).$$

In the example this might have led us to write

$$\Pr(\theta, T_1, T_2) = \Pr(\theta)\Pr(T_1|\theta)\Pr(T_2|\theta, T_1).$$

In fact the last term is just $\Pr(T_2|\theta)$ since T_2 is *conditionally independent* of T_1 *given* θ. In other words we do not draw an arrow (or *arc*) from T_1 to T_2 in figure 3.1. We can build up the joint probability as a product of one marginal probability and a sequence of conditional probabilities. The direction of the arcs denotes the order in which we are doing this and the arcs leading into a node indicate on which other unknowns we need to condition at each step. In a way, the important feature, therefore, is which possible arcs are missing. Note that we can not have a directed cycle in such a graph. The graphs are sometimes called *directed acyclic graphs* or DAGs. They are also sometimes called *influence diagrams*.

Figure 3.2 gives another example of a DAG. Here A and D are independent, B and E are conditionally independent given C and each of B, E is conditionally independent of each of A, D given C. The joint probability can be written

$$\Pr(A, B, C, D, E) = \Pr(A)\Pr(D)\Pr(C|A, D)\Pr(B|C)\Pr(E|C). \tag{3.1}$$

There is no unique graph for a group of random variables.

Suppose, in the example of figure 3.2, we wanted A to be the only node with no parents. This means reversing the direction of the arc between D and C. It is not quite as simple as this though. The joint probability is given by (3.1). By Bayes theorem,

$$\Pr(C \mid A, D) = \frac{\Pr(C \mid A)\Pr(D \mid A, C)}{\Pr(D \mid A)}$$

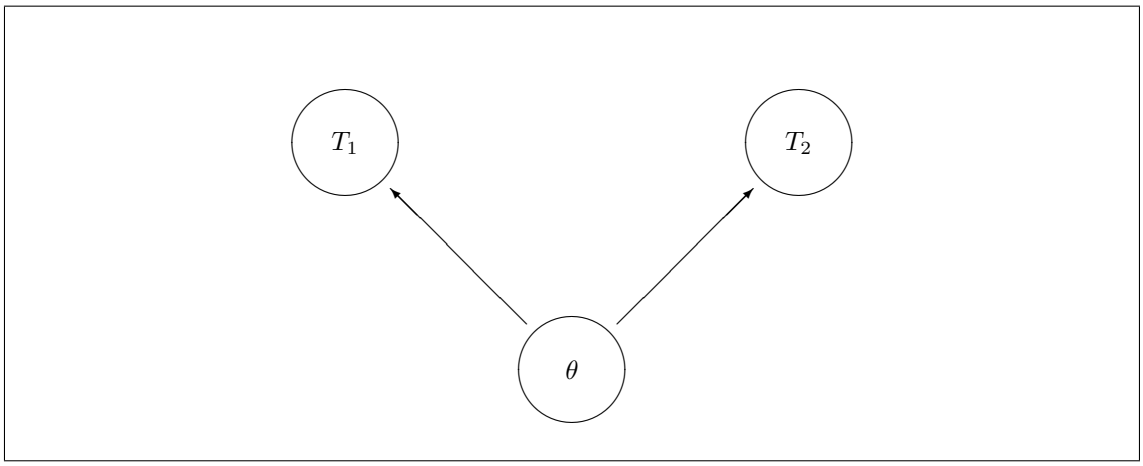

Figure 3.1: Graphical model for animals example

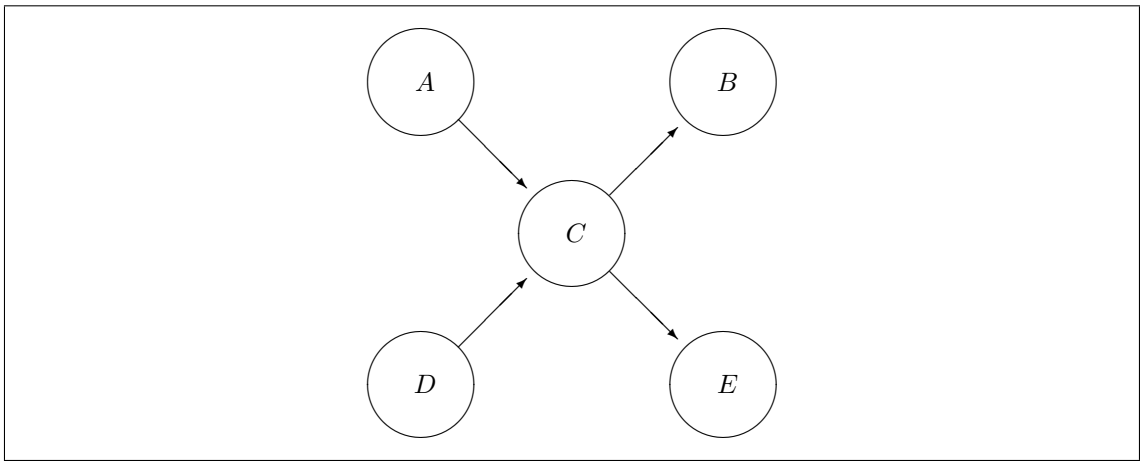

Figure 3.2: Directed acyclic graph

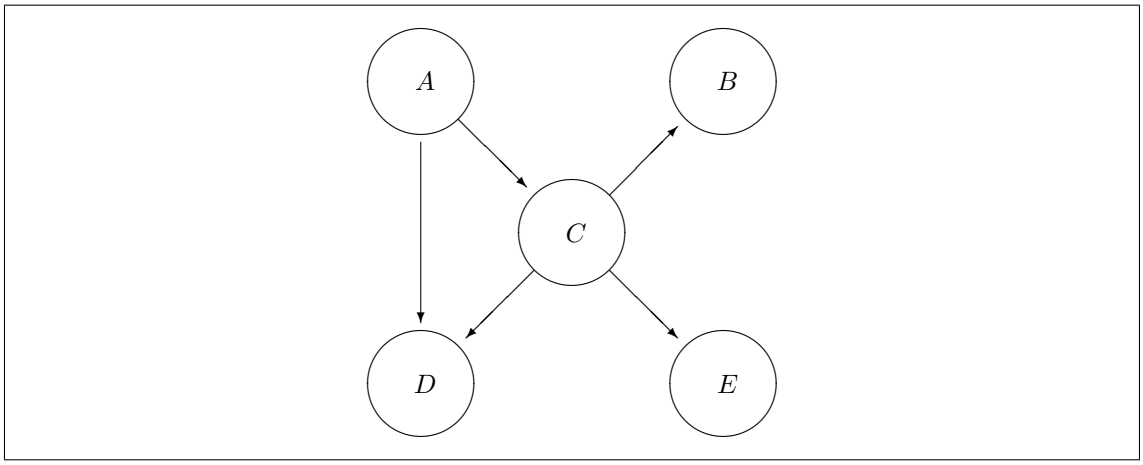

Figure 3.3: Arc reversal

but $\Pr(D \mid A) = \Pr(D)$ so

$$\Pr(D)\Pr(C \mid A, D) = \Pr(C \mid A)\Pr(D \mid A, C).$$

Hence we replace (3.1) with

$$\Pr(A, B, C, D, E) = \Pr(A)\Pr(C|A)\Pr(D|A, C)\Pr(B|C)\Pr(E|C).$$

So (unless D is conditionally independent of A given C, which would not be true in general) we need to add an arc from A to D, as in figure 3.3. The general rule is that we can reverse the direction of an arc between two nodes, N_1 and N_2, provided that

1. we do not create a directed cycle by doing so and
2. any node which is a parent of either N_1 or N_2 is made a parent of both.

See Figure 3.3.

Going back to figure 3.2, suppose we wished to eliminate C. Then we could replace (3.1) with

$$\Pr(A, B, D, E) = \Pr(A)\Pr(D)\Pr(B|A, D)\Pr(E|A, B, D)$$

which gives the diagram in figure 3.4. To see this, first note that we can *always* remove a node which has no children without having to make any other changes. This is obvious since we are just dropping a term from the end of the joint probability factorisation. So, we can arrange for the node which we want to delete to have no children by suitable arc reversals. For example, starting with figure 3.2, we could reverse the arc between B and C and then reverse the arc between C and E. This would leave us with figure 3.4. If we did the reversals in the other order we would get a slightly different result.

The procedure using arc reversals to eliminate a node will work but we might end with a graph which has more arcs than are necessary. There is a general rule which can be used which avoids this problem. (You need not memorise this rule). The general rule is as follows. If a node N is eliminated then:

- every child of N inherits all parents of N,
- every pair of the children of N is connected by an arc,
- every child that receives an arrow from another child inherits all parents of the latter.

See Pearl, Geiger and Verma (1990) p82.

Figure 3.5 shows a simple repeated measures model with three observations on each of two individuals. The model is as follows.

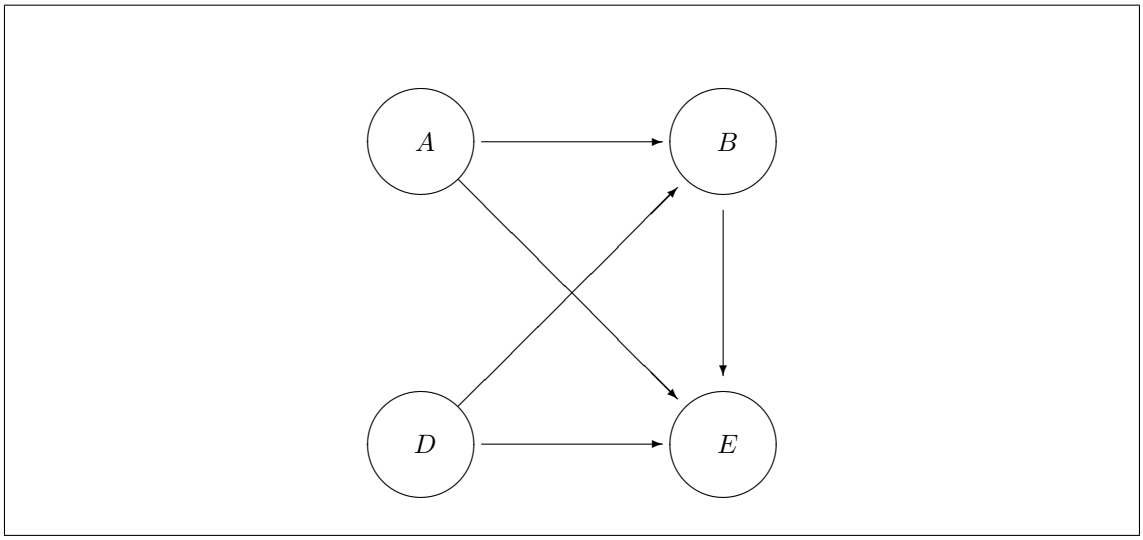

Figure 3.4: Node deletion

Suppose we have k samples of observations and observation j in sample i is

$$Y_{ij} = \theta_i + \varepsilon_{ij}$$

for $j = 1, \ldots, J$, where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $\theta_i \sim N(\mu, \sigma_\theta^2)$ (all independent).
Suppose we have independent priors for the three parameters:

$$\begin{aligned} \mu &\sim N(\mu_0, \sigma_0^2) \\ \sigma_\theta^2 &\sim \text{IG}(a_1, b_1) \\ \sigma_\varepsilon^2 &\sim \text{IG}(a_2, b_2) \end{aligned}$$

where IG stands for "inverse gamma" (i.e. $(\sigma^2)^{-1}$ has a gamma distribution).
The diagram shows the model with $k = 2$ and $J = 3$.
We will see similar models in a later lecture.

Figure 3.6 shows an example of how we might represent relationships in our prior beliefs about quantities. Here M_1, M_2, M_3 are all related because they share a common parent, U_3. Thus, if we learn something about M_1 this will affect our beliefs about M_2 and M_3. Furthermore M_1 is more strongly related to M_2 than to M_3 because M_1 and M_2 share a common parent, U_1, which is not a parent of M_3.

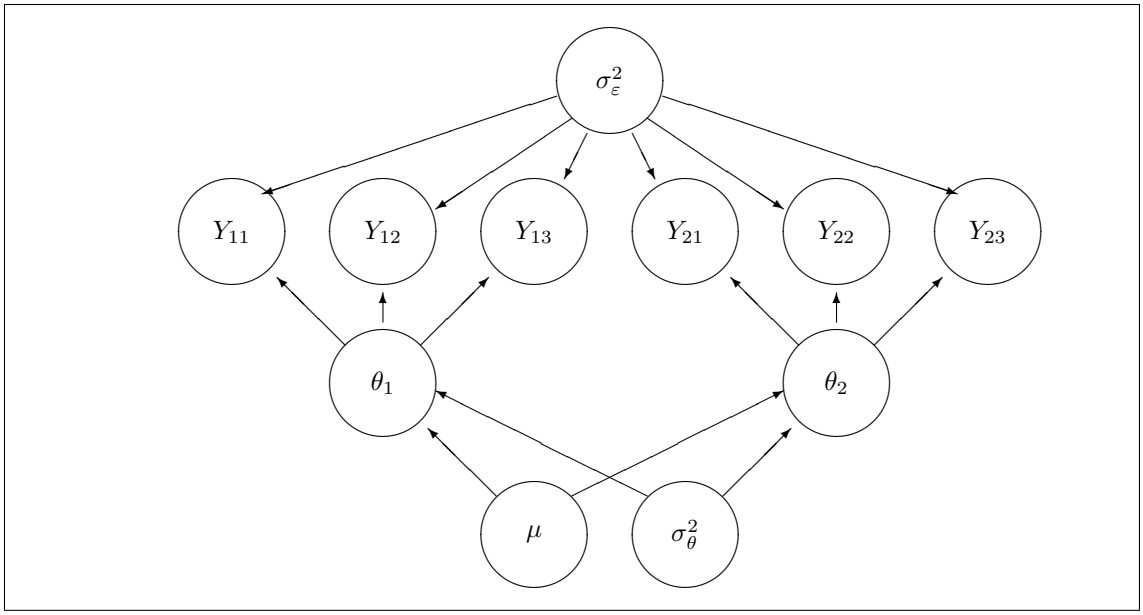

Figure 3.5: Repeated measures model

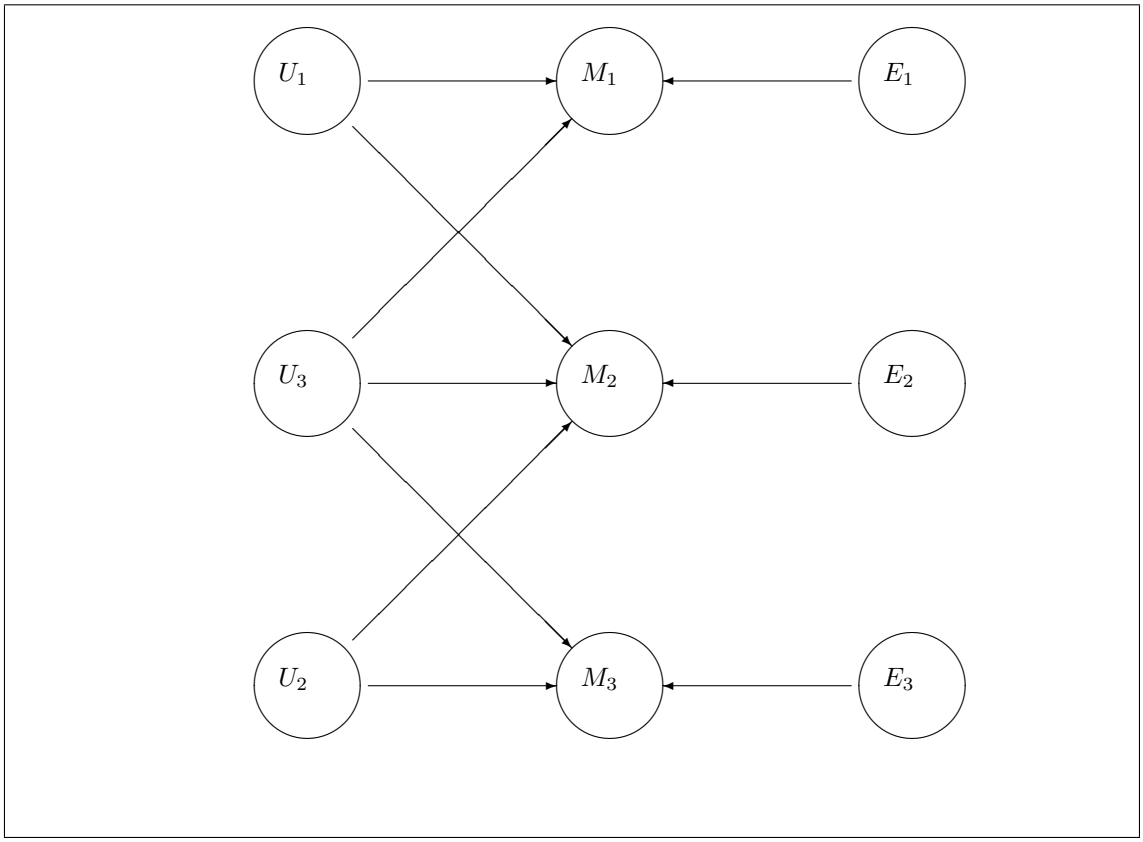

Figure 3.6: Three related quantities

3.2 Missing Data

3.2.1 Motivation

Consider a regression model with a dependent variable Y and explanatory variables $X_1, \ldots, X_p$. This could be a linear model, a generalised linear model or some other kind of model such as a survival model. In some cases the values of $X_1, \ldots, X_p$ will be deliberately chosen in a designed experiment. In other cases the data will arise from an *observational study* in which we observe the variables for each of a sample of individuals from some population. In the latter case it is possible that, for some reason, the values of one or more of the explanatory variables are missing. In fact this is not particularly unusual. Regression is not the only situation where this might be a problem. It might apply in other cases wher we make multivariate observations.

What can we do? We can not simply make inferences about the model parameters in the ordinary way when observations on some variables are missing. One possibility would be simply to delete any obervation where there are missing data. This is not satisfactory for two reasons as follow.

1. We would be losing information which could be used.
2. We might make misleading inferences as a result. It could be that the cases where data are missing are also different in some other way from the complete cases.

From the Bayesian viewpoint the missing values are simply "unknowns" in the same way as other unknowns such as model parameters. We can therefore extend our model to include a model for the variables which may be missing. If a variable X is sometimes, but not always, missing then we can (subject to certain assumptions – see below) use the cases where it is present to learn about its relationship with other variables. Therefore we can obtain a posterior distribution for the missing values. In fact we obtain a joint posterior distribution for the model parameters and the missing values. We can then "integrate out" the missing values to obtain the marginal posterior distribution of the model parameters. In effect, our inferences about model parameters are "averaged" over the distribution of possible values of the missing data. MCMC methods are well suited to handling problems of this type and it is often quite straightforward to handle missing-data problems using software such as BUGS.

We need to do two things:

1. We need to consider the nature of the "missingness" to see what it makes sense to do.
2. If it does make sense to proceed then we need a "missing data model". That is a model which shows how the variables which are sometimes missing are related to other variables.

3.2.2 Different kinds of missingness

Suppose that we make a multivariate observation $\underline{Y}$ on each of a number of individuals. In the regression example above, $\underline{Y}$ would contain both the dependent variable Y and the explanatory variable $X_1, \ldots, X_p$. For an individual i the vector of values of the variables is $\underline{y}_i = (y_{i,1}, \ldots, y_{i,J})^T$. However some of the values $y_{i,1}, \ldots, y_{i,J}$ might be missing and therefore not observed. We introduce the *inclusion indicator* $I_i = (I_{i,1}, \ldots, I_{i,J})^T$ where $I_{i,j} = 1$ if $y_{i,j}$ is observed and $I_{i,j} = 0$ if $y_{i,j}$ is missing.

We introduce (vector) parameters $\theta,\ \phi$ such that, given θ and ϕ, we can write the joint probability (density) of $\underline{y}_i, I_i$ as

$$f_{y,I}(\underline{y}_i, I_i \mid \theta, \phi) = f_y(\underline{y}_i \mid \theta) p_I(I_i \mid \underline{y}, \phi).$$

Let us consider all of the data (for all individuals) together. We write $\underline{y}$ for the complete data on all individuals, I for the inclusion indicator for all individuals, which is now a matrix, and so on. Then we write

$$f_{y,I}(\underline{y}, I \mid \theta, \phi) = f_y(\underline{y} \mid \theta) p_I(I \mid \underline{y}, \phi).$$

(The meaning of f and p has, of course, changed here).

Missing data mechanism :

The conditional distribution $p_I(I \mid \underline{y}, \phi)$ of I, given the *complete* data $\underline{y}$ and the parameters ϕ, is called the *missing data mechanism.*

Observed data and likelihood :

We divide $\underline{y}$ into $\underline{y}_{\text{obs}}$, the part which we observe, and $\underline{y}_{\text{miss}}$, the part which is missing. All that we actually observe is $\underline{y}_{\text{obs}}$ and I. The likelihood from this is therefore

$$\begin{aligned} L(\theta, \phi) &= f_{y\text{obs},I}(\underline{y}_{\text{obs}}, I \mid \theta, \phi) \\ &= \int f_y(\underline{y}_{\text{obs}}, \underline{y}_{\text{miss}} \mid \theta) p(I \mid \underline{y}_{\text{obs}}, \underline{y}_{\text{miss}}, \phi) \, d\underline{y}_{\text{miss}}. \end{aligned}$$

Missingness at random :

We say that the missing data are *missing at random* (MAR) if I is conditionally independent of the missing values given the observed values. That is

$$p(I \mid \underline{y}_{\text{obs}}, \underline{y}_{\text{miss}}, \phi) = p(I \mid \underline{y}_{\text{obs}}, \phi).$$

Then

$$\begin{aligned} L(\theta, \phi) &= f_{y\text{obs},I}(\underline{y}_{\text{obs}}, I \mid \theta, \phi) \\ &= p(I \mid \underline{y}_{\text{obs}}, \phi) \int f_y(\underline{y}_{\text{obs}}, \underline{y}_{\text{miss}} \mid \theta) \, d\underline{y}_{\text{miss}} \\ &= p(I \mid \underline{y}_{\text{obs}}, \phi) f_{y\text{obs}}(\underline{y}_{\text{obs}} \mid \theta) \end{aligned}$$

Ignorable missing-data mechanism :

Suppose that the missing data are missing at random and, in addition, the two sets of parameters $\theta,\ \phi$ are independent in our prior, so that the joint prior density is

$$f_{\theta,\phi}(\theta,\phi) = f_\theta(\theta) f_\phi(\phi).$$

Then the joint posterior density of $\theta,\ \phi$ is proportional to

$$f_\theta(\theta) f_{y\text{obs}}(\underline{y}_{\text{obs}} \mid \theta) \times f_\phi(\phi) p(I \mid \underline{y}_{\text{obs}}, \phi).$$

Therefore we can base our Bayesian inference about θ simply on the observed data $\underline{y}_{\text{obs}}$.
In this case the missing data mechanism is said to be *ignorable*.

Missingness completely at random :

Sometimes a stronger assumption than MAR is made. We say that the missing data are *missing completely at random* (MCAR) if the distribution of I does not depend on either the missing or observed values. That is

$$p(I \mid \underline{y}_{\text{obs}}, \underline{y}_{\text{miss}}, \phi) = p(I \mid \phi).$$

Notice that it is not usually necessary to assume MCAR for Bayesian inference. It is usually sufficient to have MAR - ignorable.

The MAR assumption is more plausible when we observe a large number of variables since the observed values are then more likely to convey enough information to make missingness conditionally independent of the missing values.

3.2.3 Missing data models

Consider the abrasion loss example in Practical 1 (Section 1.4.1). Suppose that some of the hardness (X_1) and tensile strength (X_2) measurements are missing.

We need a model for the joint distribution of X_1 and X_2. The existing regression model is just a model for the conditional distribution of Y given X_1 and X_2. For example, we could say

$$\begin{aligned}
X_1 \mid \mu_1, \tau_1 &\sim N(\mu_1,\ \tau_1^{-1}) \\
\mu_1 &\sim N(60,\ 400) \\
\tau_1 &\sim \text{Ga}(1,\ 100) \\
X_2 \mid x_1, \delta_1, \delta_2, \tau_2 &\sim N(\delta_1 + \delta_2[x_1 - 60],\ \tau_2^{-1})
\end{aligned}$$

$$\begin{aligned}\delta_1 &\sim N(200,\ 2500)\\ \delta_2 &\sim N(0,\ 1)\\ \tau_2 &\sim \text{Ga}(1,\ 2000)\end{aligned}$$

Notice that we have done this by giving X_1 a distribution and then giving X_2 a conditional distribution given X_1. (Of course we could have done it the other way round). There are many possibilities for the way we build a "missing data model" depending on what the variables are. For example, we might have a binary variable which we could relate to a continuous variable through a logistic regression (or we could give the continuous variable two different conditional distributions depending on the value of the binary variable).

Figure 3.7 shows a BUGS model specification in the abrasion loss example. (Note that we are not using the fully conjugate prior here). The data file would simply contain `NA` where a value is missing.

```
model

{for (i in 1:30)
     {loss[i]~dnorm(lossmean[i],tau)
      lossmean[i]<-alpha+beta[1]*(hard[i]-60)+beta[2]*(tens[i]-200)
      hard[i]~dnorm(muhard,tau.hard)
      tens[i]~dnorm(tensmean[i],tau.tens)
      tensmean[i]<-delta[1]+delta[2]*(hard[i]-60)
      }

 alpha~dnorm(150,0.000625)
 beta[1]~dnorm(0,0.0025)
 beta[2]~dnorm(0,0.0025)
 beta0<-alpha+60*beta[1]+200*beta[2]
 muhard~dnorm(60,0.0025)
 delta[1]~dnorm(200,0.0004)
 delta[2]~dnorm(0,1)

 tau.tens~dgamma(1,2000)
 tau.hard~dgamma(1,100)
 tau~dgamma(2,3200)

 }
```

Figure 3.7: BUGS model specification for abrasion loss example with missing data

3.3 Data augmentation

3.3.1 Introduction

Some models have rather complicated likelihood functions which, if handled directly, would lead to difficult calculations. Sometimes it is possible to make things much simpler by introducing extra variables, known as *auxiliary variables*, which are not observed but, which, if they were observed, would make the likelihood simpler. These auxiliary variables are then treated as if they were missing data. This is known as *data augmentation*. MCMC methods are well suited to this approach.

3.3.2 Example 1: Mixture models

In MAS3301 we looked at the use of mixture distributions as priors. We can also use mixtures as sampling distributions. The likelihood can be complicated and difficult to calculate but we can make things much simpler by introducing a group-membership variable which is unobserved. We will look at the case of mixtures in Chapter 4.

3.3.3 Example 2: Student t-model

In a normal linear model we have

$$\begin{aligned} Y_i \mid \mu_i, \sigma^2 &\sim N(\mu_i,\ \sigma^2) \\ \text{where} \quad \mu_i &= \sum_{j=1}^{J} \beta_j x_{i,j} \end{aligned}$$

Suppose instead we want to use a Student's t-distribution for the errors so

$$\frac{Y_i - \mu_i}{\sigma} \mid \mu_i, \tau \sim t_d,$$

where t_d represents the Student's t distribution on d degrees of freedom. We assume that d is chosen. A small value of d makes the error distribution "heavy-tailed."

The likelihood in this model is such that sampling for a Gibbs sampler would be difficult. However we can overcome this problem by introducing auxiliary variables X_i where

$$d\sigma^2 X_i \sim \chi^2_d.$$

Then we let

$$Y_i \mid \mu_i, X_i \sim N(\mu_i,\ X_i^{-1}).$$

Now we get the following properties.

- Y_i has the required error distribution.

  Proof: Since $d\sigma^2 X_i \sim \chi^2_d$ we have $X_i \sim \text{Ga}(d/2,\ d\sigma^2/2)$. Therefore the joint density of X_i and Y_i given μ_i and σ^2 is

$$\begin{aligned} f_{X,Y}(x_i, y_i \mid \mu_i, \sigma^2) &= (2\pi)^{-1/2} x_i^{1/2} \exp\left\{-\frac{x_i}{2}(y_i-\mu_i)^2\right\} \frac{(d\sigma^2/2)^{d/2} x_i^{d/2-1} e^{-d\sigma^2 x_i/2}}{\Gamma(d/2)} \\ &= (2\pi)^{-1/2} \frac{(d\sigma^2/2)^{d/2}}{\Gamma(d/2)} x_i^{(d+1)/2-1} e^{-x_i[d\sigma^2+(y_i-\mu_i)^2]/2} \\ &= \frac{(2\pi)^{-1/2}(d\sigma^2/2)^{d/2}\Gamma([d+1]/2)}{\Gamma(d/2)([d\sigma^2+(y_i-\mu_i)^2]/2)^{(d+1)/2}} \\ &\quad \times \frac{([d\sigma^2+(y_i-\mu_i)^2]/2)^{(d+1)/2}}{\Gamma([d+1]/2)} x_i^{(d+1)/2-1} e^{-x_i[d\sigma^2+(y_i-\mu_i)^2]/2} \end{aligned}$$

Now integrate with respect to x_i and the second term, which is a gamma density, integrates to 1. So the density of Y_i is

$$\begin{aligned} f_Y(Y_i) &= \frac{(2\pi)^{-1/2}(d\sigma^2/2)^{d/2}\Gamma([d+1]/2)}{\Gamma(d/2)([d\sigma^2+(y_i-\mu_i)^2]/2)^{(d+1)/2}} \\ &= \pi^{-1/2}d^{d/2}\sigma^d\frac{\Gamma([d+1]/2)}{\Gamma(d/2)}[d\sigma^2+(y_i-\mu_i)^2]^{-(d+1)/2} \\ &= (\pi d)^{-1/2}\sigma^{-1}\frac{\Gamma([d+1]/2)}{\Gamma(d/2)}\left[1+d^{-1}\left(\frac{y_i-\mu_i}{\sigma}\right)^2\right]^{-(d+1)/2} \end{aligned}$$

Now let $T_i = (Y_i - \mu_i)/\sigma$. Then $dt/dy = 1/\sigma$ so the density of T_i is

$$f_T(t_i) = (\pi d)^{-1/2}\frac{\Gamma([d+1]/2)}{\Gamma(d/2)}\left[1+\frac{t^2}{d}\right]^{-(d+1)/2}.$$

This is the density of a t_d distribution, as required.

- Given a multivariate normal prior for $\beta_1, \ldots, \beta_J$, the full conditional distribution of $\beta_1, \ldots, \beta_J$ (conditioning on values for $X_1, \ldots, X_J$) is also multivariate normal and therefore easy to sample.
- Given a value for σ^2 and values for $\mu_1, \ldots, \mu_n$, the full conditional distribution for each X_i is a gamma distribution and therefore easy to sample.
- Given values for $X_1, \ldots, X_n$ and a gamma prior for σ^2 (Note: σ^2 in this case, not $\tau = \sigma^{-2}$), the full conditional distribution of σ^2 is also a gamma distribution and easy to sample.

3.3.4 Example 3: Integrated moving average processes

Integrated moving average processes are commonly used as models for nonstationary time series. The integrated first order moving average process, denoted IMA (0,1,1), is especially useful.

Let the observation at time t be y_t. Let $z_t = y_t - y_{t-1}$. Then we model z_t using a first order moving average, MA(1), process

$$z_t = \varepsilon_t + \theta\varepsilon_{t-1}$$

where $\ldots, \varepsilon_{t-2}, \varepsilon_{t-1}, \varepsilon_t, \varepsilon_{t+1}, \ldots$ are independent and each has distribution $\varepsilon_j \sim N(0, \sigma^2)$, given σ^2.

This is a stationary process. For reasons of identifiability we restrict θ to $-1 \le \theta \le 1$. Given the parameters, the moments of the process are as follows.

$$\begin{aligned} \mathrm{E}(z_t) &= 0, \\ \gamma_0 = \mathrm{var}(z_t) &= \sigma^2(1+\theta^2), \\ \gamma_1 = \mathrm{covar}(z_t, z_{t-1}) &= \sigma^2\theta, \\ \gamma_k = \mathrm{covar}(z_t, z_{t-k}) &= 0 \quad (k > 1). \end{aligned}$$

(Note that the mean is zero because we have not added a nonzero "drift" into the model).

If we observe $\underline{y} = (y_1, \ldots, y_n)^T$ then, we can transform this to the equivalent observation $y_1, z_2, \ldots, z_n$ where $z_2, \ldots, z_n$ is a realisation of a MA(1) process. Given the parameters, this is an observation from a multivariate normal distribution in which the variance matrix is a function of the unknown parameters σ^2 and θ.

Things take on a simpler form if we observe that, conditional on ε_{t-1} and the model parameters, the distribution of z_t is normal with mean

$$\mathrm{E}(z_t \mid \varepsilon_{t-1}, \underline{\theta}) = \theta\varepsilon_{t-1}$$

and variance σ^2. This now looks like a straightforward normal linear regression (with no intervept because we are not fitting a nonzero mean). However, we need to know the values of ε_{t-1}. Apart

from one catch, this problem is easily solved since, if we know ε_{t-1} and z_t and the parameters, we can calculate

$$\varepsilon_t = z_t - \theta\varepsilon_{t-1}.$$

Therefore we can calculate the ε values recursively through the time series. The catch is that we need a starting value at the beginning of the series and we do not have it. That is, we do not have ε_1 which we need to calculate ε_2. The solution is to *augment* the data by including ε_1 as an auxiliary variable. Since y_1 and ε_1 are not independent and we have the observation y_1, we really ought to use this information. One way to do this is to write $y_1 = \mu + \varepsilon_1$ and specify a (normal) prior for μ.

There is another approach which is actually more popular among Bayesian statisticians (probably for historical reasons).
Suppose we write

$$\begin{aligned} y_t &= x_t + e_t, \\ x_t &= x_{t-1} + a_t, \end{aligned}$$

where $\ldots, a_{t-2}, a_{t-1}, a_t, a_{t+1}, a_{t+2}, \ldots$ and $\ldots, e_{t-2}, e_{t-1}, e_t, e_{t+1}, e_{t+2}, \ldots$ are all independent, given the parameters, and

$$\begin{aligned} a_j &\sim N(0,\ \sigma_a^2), \\ e_j &\sim N(0,\ \sigma_e^2). \end{aligned}$$

This is an example of a *dynamic linear model*.
Now, as before, let $z_t = y_t - y_{t-1}$. So

$$z_t = x_t + e_t - x_{t-1} - e_{t-1}$$

but $x_t - x_{t-1} = a_t$ so

$$z_t = e_t - e_{t-1} + a_t.$$

Now we can see that, given the parameters, the moments of the z_t process are as follows.

$$\begin{aligned} \mathrm{E}(z_t) &= 0, \\ \gamma_0 = \mathrm{var}(z_t) &= \sigma_a^2 + 2\sigma_e^2, \\ \gamma_1 = \mathrm{covar}(z_t, z_{t-1}) &= -\sigma_e^2, \\ \gamma_k = \mathrm{covar}(z_t, z_{t-k}) &= 0 \quad (k > 1). \end{aligned}$$

Therefore the two models are the same if

$$\begin{aligned} \theta\sigma^2 &= -\sigma_e^2, \\ \sigma^2(1+\theta^2) &= \sigma_a^2 + 2\sigma_e^2. \end{aligned}$$

Note that this only works provided that we are willing to restrict θ to $-1 \leq \theta \leq 0$ but this is often a reasonable assumption.

Figure 3.8 shows a possible BUGS model specification using this second approach. We can regard $x_1, \ldots, x_n$ as auxiliary data. In fact, in this model specification I have also introduced x_0 to help to construct the prior. There are different ways to construct priors for models of this sort but a method like that shown here is often used if we wish to use the model for forecasting.

```
model

{
 for (i in 1:n)
     {y[i]~dnorm(x[i],tau.y)
      }

 x0~dnorm(400,0.0001)
 x[1]~dnorm(x0,tau.x)

 for (i in 2:n)
     {x[i]~dnorm(x[i-1],tau.x)
      }

 tau.x~dgamma(1,10)
 tau.y~dgamma(1,10)

 k<- -(tau.y/tau.x+2)
 theta1<-k+sqrt(pow(k,2)-4)
 theta2<-k-sqrt(pow(k,2)-4)
 theta<-max(theta1,theta2)
 sigmasq<- -1/(theta*tau.y)
  }
```

Figure 3.8: BUGS model specification for integrated moving average IMA (0,1,1) process

3.4 Practical 3

3.4.1 Abrasion Loss

Consider the abrasion loss example in Practical 1 (Section 1.4.1). Suppose that some of the hardness (X_1) and tensile strength (X_2) measurements are missing.

1. Obtain a copy of the data file `abrasion.txt` used in Practical 1. Edit the file by changing the first two hardness values to `NA` and the third and fourth tensile strength values to `NA`. Add headings for the columns as shown below. The first few lines of the file should now look like this.

   ```
   loss  hard     tens
   372 NA 162
   206 NA 233
   175 61  NA
   154 66  NA
   136 71 231
   112 71 237
   ```

   We now have four missing values in the file.

   Save the file with the name `abmiss.txt` .

2. Obtain a copy of the BUGS model file `abmissbug.txt` from the module Web page.

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

   ```
   abmiss<-read.table("abmiss.txt",header=TRUE)
   abmissdata<-list(loss=abmiss$loss,hard=abmiss$hard,tens=abmiss$tens)
   ```

4. Create a JAGS model object.

```
model

{for (i in 1:30)
     {loss[i]~dnorm(lossmean[i],x[i])
      lossmean[i]<-alpha+beta[1]*(hard[i]-60)+beta[2]*(tens[i]-200)
      x[i]~dgamma(5,w)
      }

 alpha~dnorm(150,0.000625)
 beta[1]~dnorm(0,0.0025)
 beta[2]~dnorm(0,0.0025)
 beta0<-alpha+60*beta[1]+200*beta[2]

 w<-5*v
 v~dgamma(2,0.00125)

 }
```

Figure 3.9: BUGS model specification for abrasion loss example with Student t errors.

```
abmissjags<-jags.model("abmissbug.txt",data=abmissdata,n.chains=2)
```

5. Run the sampler for a burn-in period.

```
update(abmissjags,5000)
```

6. Start recording samples. You do not have to record everything, of course, but, in this case, it might be interesting to see the sampled values of the missing data as well as the model parameters.

```
abmisssamples<-coda.samples(abmissjags,c('beta0','beta','tau',
    'muhard','tau.hard','delta','tau.tens','hard','tens'),10000)
```

7. Look at the results in the various ways which you have seen and compare the posterior summaries with those obtained in Practical 1 when no observations were missing.

8. Try running this model again but this time with no burn-in, to see how convergence happens.

```
abmissjags<-jags.model("abmissbug.txt",data=abmissdata,n.chains=2)
abmisssamples<-coda.samples(abmissjags,c('beta0','beta','tau',
    'muhard','tau.hard','delta','tau.tens','hard','tens'),10000)
par(ask=TRUE)
traceplot(abmisssamples)
```

Etc.

3.4.2 Abrasion loss with t errors

Let us analyse the abrasion loss data again, this time with no missing data but with Student t errors.

1. Create a new model file, perhaps by editing `abmissbug.txt`. Save the file as `abtbug.txt`. The file should look like Figure 3.9. We set the degrees of freedom to 10. We will use data augmentation (although, in fact, `rjags` is probably clever enough to handle the problem even without this).

460	457	452	459	462	459	463	479	493	490	492	498	499	497	496	490
489	478	487	491	487	482	479	478	479	477	479	475	479	476	476	478
479	477	476	475	475	473	474	474	474	465	466	467	471	471	467	473
481	488	490	489	489	485	491	492	494	499	498	500	497	494	495	500
504	513	511	514	510	509	515	519	523	519	523	531	547	551	547	541
545	549	545	549	547	543	540	539	532	517	527	540	542	538	541	541
547	553	559	557	557	560	571	571	569	575	580	584	585	590	599	603
599	596	585	587	585	581	583	592	592	596	596	595	598	598	595	595
592	588	582	576	578	589	585	580	579	584	581	581	577	577	578	580
586	583	581	576	571	575	575	573	577	582	584	579	572	577	571	560

Table 3.1: One hundred and sixty consecutive daily IBM common stock closing prices. The data are to be read along the rows.

2. Create a data file, called `abrasion.txt`, like `abmiss.txt` but with none of the observations missing.

3. Try the analysis.

```
abrasion<-read.table("abrasion.txt",header=TRUE)
abdata<-list(loss=abrasion$loss,hard=abrasion$hard,tens=abrasion$tens)
abtjags<-jags.model("abtbug.txt",data=abdata,n.chains=2)
update(abtjags,(5000)
abtsamples<-coda.samples(abmissjag,c('beta0','beta','v'),10000)
par(ask=TRUE)
traceplot(abtsamples)
```

   Etc.

4. Compare the results with the results in Practical 1. What do you think is the effect of using t rather than normal errors?

3.4.3 IBM Stock Prices

Table 3.1 shows 160 consecutive daily IBM common stock closing prices. The data may be obtained from the Module Web Page. They were obtained from Box and Jenkins (1976). Box and Jenkins suggest fitting an IMA (0,1,1) model to these data.

1. Use the model file shown in Figure 3.8 to analyse these data. You can obtain the model file from the Module Web Page.

2. We can calculate forecasts in a straightforward way. Edit the data file. Change `n=160` to `n=164`. Add `NA` four times at the end of the list of `y` values, separated by commas. Repeat the analysis but this time you can record the values of the "missing" Y values, by monitoring `y`, and thus obtain a forecast distribution.

3.5 Exercise

Table 3.2 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 $\text{Bin}(n_{j,k},\ p_{j,k})$ 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, $\Pr(p_0 < 0.02) = \Pr(p_0 > 0.10) = 0.025$. Find the mean and variance of our normal prior distribution for α_0.

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

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

   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. Find the values of $V_{\alpha,0}$ and $V_{\alpha,1}$.

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.

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

   Here $\beta_1, \ldots, \beta_4$ are conditionally independent given $\bar{\beta}$. We choose to make $V_{\beta,0} = V_{\beta,1}$ and $V_{\beta,0} + V_{\beta,1} = 0.0004$. Find the values of $V_{\beta,0}$ and $V_{\beta,1}$. The value of m_β is 0.0.

4. Construct a suitable BRugs model file. Hint: You can use a construction such as `alpha[group[i]]` to denote α_j where observation i belongs to group j.

5. The data are available from the Module Web Page 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.

6. Present summaries of the inference, including posterior means and standard deviations of the parameters.

7. Find the posterior mean and standard deviation 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. Plot the psoterior probability density function of this quantity and comment.

	Area 1				Area 2			
	Total undergoing surgery		Number dying		Total undergoing surgery		Number dying	
Age	Males	Females	Males	Females	Males	Females	Males	Females
5-14	4272	3911	9	11	1739	1758	5	2
15-24	2835	2989	23	5	1233	1244	14	1
25-34	2785	2606	19	8	989	1004	8	3
35-44	1930	1886	16	15	897	922	9	13
45-54	1497	1524	59	40	921	961	28	15
55-64	960	1013	101	52	686	739	68	37
65-75	652	855	185	118	611	784	159	73
76-83	186	287	97	108	189	290	86	88

Table 3.2: Deaths following surgery in two areas of the USA

3.6 Problems 4

Solutions to all questions are to be submitted in the Homework Letterbox no later than 4.00pm on Wednesday December 12th. Please note that you should give some attention to the presentation of your work. Describe the data, model, prior etc. and explain what you have done. Comment on your conclusions. A listing of the output from a R session with one or two things written on it will not get a very good mark on its own.

In questions 2 and 3, each student is given different data. For this purpose each student is given a reference number according to the table below. Please use the correct data and write your reference number on your work.

Reference numbers

Problems

1. Full conditional distribution

   Certain components are manufactured in batches. Each batch contains n components. The components in N batches are then tested and some are found to be defective. Let the number of defective components in batch i be x_i. We suppose that, given the value of π_i, where $0 < \pi_i < 1$, the value of x_i is an observation from a binomial distribution $X_i \sim \text{Bin}(n, \pi_i)$ and X_i and X_j are independent given π_i and π_j when $i \neq j$. Let $\eta_i = \log_e\{\pi_i/(1-\pi_i)\}$. We suppose that, given the values of μ and τ, η_i is an observation from the normal $N(\mu, \tau^{-1})$ distribution and η_i and η_j are independent, when $i \neq j$, given the values of μ and τ. Finally we have independent prior distributions for μ and τ with μ having a normal prior, $\mu \sim N(m, v)$, and τ having a gamma prior, $\tau \sim \text{Ga}(a, b)$.

   Find a function proportional to the density of the full conditional distribution (fcd) of η_i, that is the distribution of η_i given x_i and values for μ and τ.

   (10 marks)

2. Piston rings

Four compressors are located in the same building. Each has three "legs". The compressors are of the same design and are oriented the same way. The three legs of each are labelled "North", "Centre" and "South." Over a certain period of time the number of failures of piston rings in each leg of each compressor is counted. These numbers are your data.

The model is as follows. Let the number of failures in leg i of compressor j be $y_{i,j}$ (where $i = 1$ for North, $i = 2$ for Centre and $i = 3$ for South). Given the value of a quantity $\lambda_{i,j} > 0$, we assume that $y_{i,j}$ is an observation from a Poisson distribution $Y_{i,j} \sim \text{Po}(\lambda_{i,j})$, with $Y_{i,j}$ independent of $Y_{i',j'}$ unless $(i,j) = (i',j')$, given the values of $\lambda_{i,j}$ and $\lambda_{i',j'}$.

The prior distribution is as follows. Let $\eta_{i,j} = \log_e(\lambda_{i,j})$. Then

$$\eta_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$$

where, $\alpha_1, \ldots, \alpha_3, \beta_1, \ldots, \beta_4, \gamma_{1,1}, \ldots, \gamma_{3,4}$ and μ are mutually independent and

$$\begin{aligned}
\mu &\sim N(3,4), \\
\alpha_i &\sim N(0,1), \qquad i = 1, \ldots, 3, \\
\beta_j &\sim N(0,1), \qquad j = 1, \ldots, 4, \\
\gamma_{i,j} &\sim N(0, 0.25), \qquad i = 1, \ldots, 3,\ j = 1, \ldots, 4.
\end{aligned}$$

Data.

To obtain your data, first install the R function `pistonread`. This function may be obtained from the Module Web Page, under "Data", or directly from

`http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/pistonreadR.txt`

or simply by typing the following into R.

```
pistonread<-function(refno)
{data<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/pistondata.txt")
 out<-cbind(data[,1],data[,2],data[,refno])
 write.table(format(out),row.names=FALSE,col.names=FALSE,quote=FALSE,file="mypistondata.txt")
 }
```

Then, use the function with your reference number as the argument. For example, if your reference number is 20, type

```
pistonread(20)
```

You will then have a file (in your working directory) called `mypistondata.txt`.There will be three columns of data, as follows.

- The leg numbers (1 for North *etc*) are in column 1.
- The compressor numbers are in column 2.
- Your failure numbers are in column 3.
- Use MCMC to take samples from the posterior distribution of the unknowns in the model.

(5 marks)

- Display your results appropriately.

(4 marks)

- Explain your method and show your BUGS model specification and the commands which you have used.

(4 marks)

- Show how you have checked convergence.

(3 marks)

- Give summaries of the posterior distributions of the model unknowns. In particular, compare the failure rates in the twelve legs using the posterior distribution. What can you conclude?

(4 marks)

3. Fraud

Banks and credit card companies attempt to detect fraud by looking for unusual observations in the withdrawal data for customers. This potentially involves quite complicated models. The model in this question is a somewhat simplified version but the principal is the same.

You will each be supplied with data for five customers. For each of these customers you will be given the total withdrawals from the customer's account for each of twenty weeks. The value for customer j in week i is $y_{i,j}$.

For each customer in each week there is a small probability π that a fraud takes place. We therefore use a mixture model with two components. The component indicator for customer j in week i is $c_{i,j}$.

If $c_{i,j} = 1$ then a fraud against customer j takes place in week i.

If $c_{i,j} = 2$ then no fraud takes place against customer j in week i.

We assume that, given the model parameters, $c_{i,j}$ is independent of $c_{i',j'}$ for $(i,j) \neq (i',j')$. Given π, we have $\Pr(c_{i,j} = 1) = \pi$. Our prior distribution for π is $\text{Beta}(1, 99)$.

If $c_{i,j} = 1$ then $y_{i,j} \sim \text{Ga}(2,\ 0.0002)$. If $c_{i,j} = 2$ then, given $\alpha,\ \beta_j$, we have $y_{i,j} \sim \text{Ga}(\alpha, \beta_j)$. We assume that $y_{i,j}$ is independent of $y_{i',j'}$ for $(i,j) \neq (i',j')$, given $\alpha,\ \beta_j$ and $\beta_{j'}$. Our prior distribution for α is $\text{Ga}(2,\ 0.5)$.

Let $\beta_j = \alpha/\lambda_j$ and $\lambda_j = \exp(\mu_j)$. Given $\mu_0,\ \tau$, we have $\mu_j \sim N(\mu_0, \tau^{-1})$ with μ_j independent of $\mu_{j'}$ for $j \neq j'$. Our prior distribution for μ_0 is $\mu_0 \sim N(5.3, 1.4)$. Our prior distribution for τ is $\text{Ga}(3, 4)$.

Unless otherwise stated, the prior distributions are independent.

Data. To obtain your data, first install the R function `fraudread`. This function may be obtained from the Module Web Page, under "Data", or directly from

```
http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/fraudreadR.txt
```

or simply by typing the following into R.

```
fraudread<-function(refno)
{data<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/frauddata.txt")
 no1<-5*(refno-11)+1
 no5<-no1+4
 out<-data[,no1:no5]
 write.table(format(out),row.names=FALSE,col.names=FALSE,quote=FALSE,file="myfrauddata.txt")
 }
```

Then, use the function with your reference number as the argument. For example, if your reference number is 20, type

```
fraudread(20)
```

You will then have a file (in your working directory) called `myfrauddata.txt`.There will be five columns of data, one for each customer, and twenty rows, one for each week. The file will then be ready to use as a data file with BRugs.

- Use MCMC to find the posterior means for $p_{i,j} = 2 - c_{i,j}$ and hence find the posterior probabilities of fraud for each customer in each week and identify any cases where fraud is likely to have occurred.

  (5 marks)

- Display your results appropriately.

  (4 marks)

- Explain your method and show your BUGS model specification and the commands which you have used.

  (4 marks)

- Show how you have checked convergence.

  (3 marks)

- Give summaries of the posterior distributions of the model parameters.

  (4 marks)

Chapter 4

Mixture Models

4.1 Mixtures

4.1.1 Finite mixtures as sampling distributions

In MAS3301 we looked at the use of mixture distributions as priors. We can also use mixtures as sampling distributions.

There are two reasons why we might want to do this:

1. We might believe that there really are two or more sub-populations and it makes sense to represent each by a component of the mixture. For example, the amount of a compound found in blood samples taken from animals might depend on whether or not the animal carries a particular infection. There are thus two sub-populations, one of infected animals and one of non-infected animals. We might not know which animals are infected but it might make sense to allow for these two sub-populations in a model. The distribution of the amount of the substance might be bimodal.

2. Even when there is no "physical" interpretation of the mixture components, using a mixture distribution allows more flexibility in the sampling model. We can relax the assumption that the data are "normally distributed", for example.

Consider a simple two-component mixture model. Our sampling model for observation Y_i has pdf

$$f(y_i;\ \pi, \theta_1, \theta_2) = \pi f_1(y_i;\theta_1) + (1-\pi) f_2(y_i;\theta_2).$$

Here $f_j(y;\theta_j)$ is the pdf for component j and depends on parameters θ_j. The component membership probabilities are π and $1-\pi$, with $0 \leq \pi \leq 1$.

Suppose that we have n independent (given the parameters) observations $y_1, \ldots, y_n$. The likelihood is

$$L = \prod_{i=1}^{n} \left\{ \pi f_1(y_i;\theta_1) + (1-\pi) f_2(y_i;\theta_2) \right\}. \tag{4.1}$$

This has a rather complicated form. For example, it is a polynomial of degree n in π.

More generally we could have J components with

$$f(y_i;\ \underline{\pi}, \Theta) = \sum_{j=1}^{J} \pi_j f_j(y_i;\theta_j), \tag{4.2}$$

where $\sum_{j=1}^{J} \pi_j = 1$ and $\pi_j \geq 0$ for $j = 1, \ldots, J$. In this case the likelihood is

$$L = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} \pi_j f_j(y_i;\theta_j) \right\}. \tag{4.3}$$

This could be very complicated.

We can make things much simpler by introducing a group-membership variable which is unobserved. The values form auxiliary data so this is an example of data augmentation.

We introduce, for observation i, an auxiliary variable c_i, which can take the values $1, \ldots, J$. Then, given that $c_i = j$, the conditional pdf for observation i is simply $f_j(y_i;\ \theta_j)$. The corresponding conditional likelihood is then just

$$L_c = \prod_{i=1}^{n} \pi_{c_i} f_{c_i}(y_i;\ \theta_{c_i}).$$

Now we give c_i a multinomial (or "categorical") distribution, in which $\Pr(c_i = j) = \pi_j$. We give the parameters $\underline{\pi} = (\pi_1, \ldots, \pi_J)^T$ and $\Theta = \{\theta_1, \ldots, \theta_J\}$ a suitable prior distribution. Then, by "integrating out", i.e. "averaging over", $c_1, \ldots, c_n$, we obtain the correct posterior distribution.

The joint probability (density) that $c_i = j$ and $Y_i = y_i$ is

$$f(y_i, c_i = j;\ \underline{\pi}, \Theta) = \pi_j f_j(y_i;\ \theta_j).$$

To find the marginal probability density of y_i we sum over j and obtain (4.2) as required.

4.1.2 MCMC and label-switching

MCMC

Once we have the model set up with the auxiliary variables $c_1, \ldots, c_n$ as above, we have a prior distribution with density $f_0(\Theta, \underline{\pi})$ for the parameters and we have initial values for the unknowns, $\Theta,\ \underline{\pi},\ c_1, \ldots, c_n$, then we can proceed with MCMC as follows.

1. Sample a new value for Θ.

The fcd density is proportional to

$$f_0(\Theta, \underline{\pi}) \prod_{j=1}^{J} L_{c,j}$$

where

$$L_{c,j} = \prod_{i \in C_j} f_j(y_i;\ \theta_j)$$

and $i \in C_j$ if $c_i = j$. That is C_j is the set of observations currently assigned to component j. We might well have $f_0(\Theta, \underline{\pi}) = f_{0,\theta}(\Theta) f_{0,\pi}(\underline{\pi})$ in which case the fcd density is proportional to

$$f_0(\Theta) \prod_{j=1}^{J} L_{c,j}$$

2. Sample a new value for $\underline{\pi}$.

The fcd density is proportional to

$$f_0(\Theta, \underline{\pi}) \prod_{j=1}^{J} \pi_j^{n_j}$$

where n_j is the number of observations currently assigned to component j. If $f_0(\Theta, \underline{\pi}) = f_{0,\theta}(\Theta) f_{0,\pi}(\underline{\pi})$ then the fcd density is proportional to

$$f_{0,\pi}(\underline{\pi}) \prod_{j=1}^{J} \pi_j^{n_j}.$$

A popular choice for $f_{0,\pi}(\underline{\pi})$ would be a Dirichlet density. In this case the fcd is also a Dirichlet distribution. Sampling from a Dirichlet distribution is quite easy.

3. Sample a new value for each of $c_1, \ldots, c_n$.

The fcd is a categorical distribution with

$$\Pr(c_i = j) \propto \pi_j f_j(y_i;\ \theta_j).$$

4. Repeat.

Label-switching

Consider the likelihood (4.1).
Suppose that both component distributions are of the same family so that the likelihood is

$$L = \prod_{i=1}^{n} \{\pi f_y(y_i; \theta_1) + (1-\pi) f_y(y_i; \theta_2)\}.$$

Suppose that we "switch the labels" and write

$$\tilde{L} = \prod_{i=1}^{n} \left\{\tilde{\pi} f_y(y_i; \tilde{\theta}_1) + (1-\tilde{\pi}) f_y(y_i; \tilde{\theta}_2)\right\}$$

where $\tilde{\pi} = 1 - \pi$, $\tilde{\theta}_1 = \theta_2$ and $\tilde{\theta}_2 = \theta_1$.

Clearly $L = \tilde{L}$. The likelihood is therefore bimodal and, in fact, the modes match each other. If the prior does not strongly favour one mode over the other then the posterior distribution will also be bimodal.

In the more general case of (4.3) we can also permute the component labels and get the same likelihood (provided that the distributions are all of the same family). This time the posterior will be multimodal unless the prior strongly favours one mode.

Unless we do something about this, it can cause difficulties in MCMC sampling using the data-augmentation method. If the posterior is multimodal then, eventually, the sampler will jump from one mode to another. The auxiliary variables c_i will suddenly change values so that observations move from one component to another and the parameters "go with them." This might only happen after thousands of iterations. Therefore we might need a very large number of iterations before the sampler has stayed in each mode the correct proportion of the time.

Clearly this behaviour is undesirable. If θ_j is a scalar parameter we can (usually) avoid the problem by imposing an order constraint on the parameters. That is by requiring that $\theta_1 < \theta_2 < \cdots < \theta_J$.

I say "usually" because we can encounter another problem It may be that our mixture model has J components but the data, through the likelihood, suggest only $J-1$ components. Then we might encounter switching between different possibilities for which label is the absent component. There are more advanced methods, beyond the scope of this module, which can deal with this problem.

When θ_j is a vector parameter we may need more ingenuity to devise suitable constraints.

4.1.3 Multivariate mixtures

It is, of course, possible to make a mixture model where the observation $\underline{y}$ is multivariate. For example, we might make several measurements on each of a sample of birds belonging to one species with the idea that there might be two or more subspecies. In two dimensions we might expect a plot of observations y_1 against y_2 to reveal "clusters" of observations.

4.1.4 Continuous mixtures

As well as the finite mixtures described above it is possible to have a mixture model with an infinite number of components. It is also possible to have a *continuous mixture*. In a continuous mixture model, instead of (4.2), we have, for example,

$$f(y_i) = \int_\Omega f_\theta(\theta) f_y(y_i; \theta, \lambda_i)\ d\theta. \tag{4.4}$$

Here θ is a parameter with a continuous distribution specified by the *mixing density* $f_\theta(\theta)$. The range of values of θ is denoted by Ω. There may be other parameters which do not vary in this way and these are denoted by λ_i.

We saw an example of this in Section 3.3.3 where we used Student-t errors in a regression. The model was

$$\begin{aligned} Y_i \mid \mu_i, X_i &\sim N(\mu_i,\ X_i^{-1}), \\ d\sigma^2 X_i &\sim \chi^2_d. \end{aligned}$$

Here μ_i corresponds to λ_i in (4.4) and X corresponds to θ in (4.4). The mixing density is that of a scaled χ^2 distribution and $f_y(y_i; \theta, \lambda_i)$ in (4.4) corresponds to $\phi(X_i^{1/2}[y_i - \mu_i])$ where $\phi()$ is the standard normal pdf.

4.2 Mixture Examples

4.2.1 "Old Faithful"

Table 4.1 shows 299 successive waiting times, in minutes, between the starts of eruptions of the "Old Faithful" geyser in the Yellowstone National Park, Wyoming, USA. The data are taken from Azzalini and Bowman (1990).

Figure 4.1 shows histograms of the data and the logs of the data. In each case we appear to see two distinct modes. However the human brain is very good at spotting patterns, even when they are not there. The evidence in the data migh not be as strong as we imagine. Each of the two-component mixture models below has five parameters, compared to two parameters for a simple normal or gamma model. The likelihood might not distinguish very strongly between all possible values of these five parameters. Therefore careful choice of a prior distribution might be important. If we really believe that there are two sub-populations then our prior may need to reflect this.

Normal mixture

Let us try using a two-component normal mixture model for the log intervals. So

$$\begin{aligned}
\Pr(c_i = 1) &= \pi \\
\Pr(c_i = 2) &= 1 - \pi \\
\pi &\sim \text{Beta}(a_\pi, b_\pi) \\
y_i \mid \mu_j, \tau_j, c_i = j &\sim N(\mu_j,\ \tau_j^{-1}) \\
\mu_j \mid \mu_0 &\sim N(\mu_0 + \delta_j,\ \tau_\mu^{-1}) \\
\mu_0 &\sim N(M_\mu,\ V_\mu) \\
\tau_j &\sim \text{Ga}(a_\tau, b_\tau)
\end{aligned}$$

Notice that we have given $\mu_1,\ \mu_2$ a "hierarchical prior." Each depends on μ_0 which then has a prior of its own. In order to avoid label switching we can impose the restriction $\mu_1 < \mu_2$. We also push the conditional prior means of μ_1, μ_2 apart by making them $\mu_0 + \delta_1$ and $\mu_0 + \delta_2$ respectively, where $\delta_1 = -\delta$ and $\delta_2 = \delta$.

We could also use a hierarchical prior for τ_1 and τ_2 although this is not quite as straightforward with gamma distributions as it is with normal distributions. I have just given them independent priors here. There is no need to impose an order constraint on $\tau_1,\ \tau_2$.

The specification of the prior is completed by giving numerical values to $a_\pi, b_\pi, a_\tau, b_\tau, M_\mu, V_\mu, \tau_\mu, \delta$. We will use the following values.

$$a_\pi = 3, \quad b_\pi = 3, \quad a_\tau = 4, \quad b_\tau = 0.04,$$

$$M_\mu = 4.0 \approx \log(60), \quad V_\mu = 0.30 \approx (\log(3)/2)^2, \quad \tau_\mu = 3.3 \approx (\log(3)/2)^{-2}, \quad \delta = 0.2.$$

Figure 4.2 shows a BUGS model specification for this example.

80	71	57	80	75	77	60	86	77	56	81	50	89	54	90
73	60	83	65	82	84	54	85	58	79	57	88	68	76	78
74	85	75	65	76	58	91	50	87	48	93	54	86	53	78
52	83	60	87	49	80	60	92	43	89	60	84	69	74	71
108	50	77	57	80	61	82	48	81	73	62	79	54	80	73
81	62	81	71	79	81	74	59	81	66	87	53	80	50	87
51	82	58	81	49	92	50	88	62	93	56	89	51	79	58
82	52	88	52	78	69	75	77	53	80	55	87	53	85	61
93	54	76	80	81	59	86	78	71	77	76	94	75	50	83
82	72	77	75	65	79	72	78	77	79	75	78	64	80	49
88	54	85	51	96	50	80	78	81	72	75	78	87	69	55
83	49	82	57	84	57	84	73	78	57	79	57	90	62	87
78	52	98	48	78	79	65	84	50	83	60	80	50	88	50
84	74	76	65	89	49	88	51	78	85	65	75	77	69	92
68	87	61	81	55	93	53	84	70	73	93	50	87	77	74
72	82	74	80	49	91	53	86	49	79	89	87	76	59	80
89	45	93	72	71	54	79	74	65	78	57	87	72	84	47
84	57	87	68	86	75	73	53	82	93	77	54	96	48	89
63	84	76	62	83	50	85	78	78	81	78	76	74	81	66
84	48	93	47	87	51	78	54	87	52	85	58	88	79	

Table 4.1: Waiting times, in minutes, between eruptions of the "Old Faithful" geyser. Data are to be read along the rows.

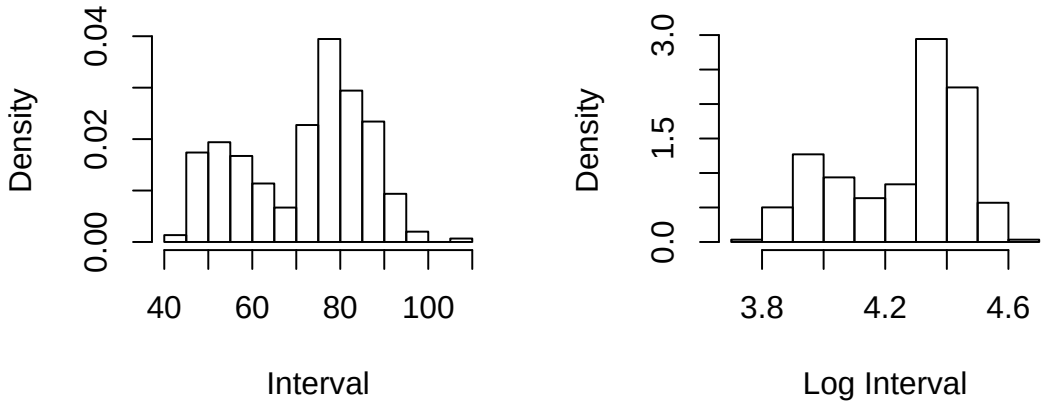

Figure 4.1: Histograms of time intervals between eruptions of "Old Faithful" and logs of the intervals.

```
model

{for (i in 1:n)
   {c[i]~dcat(q[])
    y[i]~dnorm(mu[c[i]],tau[c[i]])
    }

 for (j in 1:2)
   {tau[j]~dgamma(4,0.04)
    }

 mumean[1]<-mu0-0.2
 mumean[2]<-mu0+0.2
 for (j in 1:2)
   {mudash[j]~dnorm(mumean[j],3.3)
    }
 mu[1:2]<-sort(mudash) # This imposes the order constraint.


 mu0~dnorm(4.0,p.mu)
 p.mu<-1/0.3

 pi~dbeta(3,3)
 q[1]<-pi
 q[2]<-1-pi
 }
```

Figure 4.2: BUGS model specification for "Old Faithful" normal mixture model.

Gamma mixture

As an alternative to the normal mixture for the log intervals, which is, of course, equivalent to a lognormal mixture for the intervals, we could try a gamma mixture for the intervals themselves.

$$
\begin{aligned}
\Pr(c_i = 1) &= \pi \\
\Pr(c_i = 2) &= 1 - \pi \\
\pi &\sim \text{Beta}(a_\pi, b_\pi) \\
t_i \mid \alpha_j, \beta_j, c_i = j &\sim \text{Ga}(\alpha_j,\ \beta_j) \\
\beta_j &= \alpha_j / \lambda_j \\
\lambda_j &= \exp(\mu_j) \\
\mu_j \mid \mu_0 &\sim N(\mu_0 + \delta_j,\ \tau_\mu^{-1}) \\
\mu_0 &\sim N(M_\mu,\ V_\mu) \\
\alpha_j &\sim \text{Ga}(a_\alpha, b_\alpha)
\end{aligned}
$$

Since the mean of a gamma(α_j, β_j) distribution is α_j/β_j and we set $\beta_j = \alpha_j/\lambda_j$, the mean interval, in component j, is λ_j. We then treat $\mu_j = \log(\lambda_j)$ in the same way as we treated μ_j in the lognormal mixture. Of course the log of the mean is not the same as the mean of the logs but, in

```
model

{for (i in 1:n)
   {c[i]<-dcat(q[])
    t[i]~dgamma(alpha[c[i]],beta[c[i]])
    }

 for (j in 1:2)
   {alpha[j]~dgamma(3,0.1)
    beta[j]<-alpha[j]/lambda[j]
    lambda[j]<-exp(mu[j])
    }


 mumean[1]<-mu0-0.2
 mumean[2]<-mu0+0.2
 for (j in 1:2)
   {mudash[j]~dnorm(mumean[j],3.3)
    }
 mu[1:2]<-sort(mudash) # This imposes the order constraint.

 mu0~dnorm(4.0,p.mu)
 p.mu<-1/0.3

 pi~dbeta(3,3)
 q[1]<-pi
 q[2]<-1-pi
 }
```

Figure 4.3: BUGS model specification for "Old Faithful" gamma mixture model.

this case, this difference has little effect. (To avoid this slight discrepancy we would have to make λ_j the median rather than the mean but this is not convenient with a gamma distribution).

I have not used a hierarchical prior for α_1, α_2. I have just given them independent priors here. There is no need to impose an order constraint on $\alpha_1,\ \alpha_2$.

We will use the following values to complete the prior specification.

$$a_\pi = 3, \quad b_\pi = 3, \quad a_\alpha = 3, \quad b_\alpha = 0.1,$$

$$M_\mu = 4.0 \approx \log(60), \quad V_\mu = 0.30 \approx (\log(3)/2)^2, \quad \tau_\mu = 3.3 \approx (\log(3)/2)^{-2}, \quad \delta = 0.2.$$

Figure 4.3 shows a BUGS model specification for this example.

4.2.2 Road vehicle headways

Cowburn (2003) describes the use of mixture models for the time gaps, or "headways", between vehicles passing along a road. See also Cowburn and Farrow (2007). The idea is that headways fall naturally into one of two sub-populations:

1. Headways where the following vehicle is not impeded by the vehicle in front.
2. "Congested" headways where the following vehicle is impeded by the vehicle in front.

Cowburn proposed that non-congested headways would follows an exponential distribution, that is a $\text{Ga}(1, \beta_1)$ distribution, and congested headways would follow a $\text{Ga}(\alpha_2, \beta_2)$ distribution with $\alpha_2 > 1$. (A number of other mixture models have been proposed in the highway engineering literature). Successive headways are independent (given the model parameters) in this version of the model. We will see a version where this is not the case in the next lecture.

```
model

{

 for (i in 1:N)
     {c[i]~dcat(q[])
      t[i]~dgamma(alpha[c[i]],beta[c[i]])
      }


 alpha[1]<-1
 alpha[2]<-1+aa
 aa~dgamma(1,0.5)
 rhodash[1]~dgamma(2,8)
 rhodash[2]~dgamma(2,4)
 rho[1:2]<-sort(rhodash)
 for (j in 1:2)
     {beta[j]<-alpha[j]*rho[j]
      }

 pi~dbeta(1,2)
 q[1]<-pi
 q[2]<-1-pi

 }
```

Figure 4.4: BUGS specification for independent headways model.

A BUGS model specification is given in Figure 4.4. The constraint that $\alpha_2 > 1$ is imposed by letting $\alpha_2 = 1 + A$ where $A \sim \text{Ga}(a_A, b_A)$. In the BUGS code A is represented by `aa`. We have $a_A = 2$ and $b_A = 8$. The mean headway in component j is $\mu_j = \rho_j^{-1} = \alpha j/\beta_j$ where $\alpha_1 = 1$. We set $\beta_j = \alpha_j \rho_j$. We ensure that $\mu_1 > \mu_2$ by ensuring that $\rho_1 < \rho_2$. The headways are recorded in seconds.

4.3 Hidden Markov Models

4.3.1 Introduction

In the mixture models above, the unobserved (or *latent*) component memberships c_i are independent of each other, given the model parameters. Sometimes, when the data have a natural ordering, as in a time series, we may wish to allow the component memberships to depend on each other.

Figure 4.5 shows the logarithms of the time intervals between eruptions of "Old Faithful". The i^{th} log interval y_i is plotted against the preceding log interval y_{i-1}. Clearly successive intervals are not independent. One way to model this might be to suppose that there are "short intervals" and "long intervals" and that a short interval is always followed by a long interval but a long interval may be followed by either a short interval or a long interval. Thus we could model the sequence $c_1, \ldots, c_n$ using a two-state Markov chain with the following transition matrix, where $q_{j,k} = \Pr(c_i = j \mid c_{i-1} = k)$.

$$\begin{pmatrix} q_{1,1} & q_{1,2} \\ q_{2,1} & q_{2,2} \end{pmatrix} = \begin{pmatrix} 0 & \pi \\ 1 & 1-\pi \end{pmatrix}. \tag{4.5}$$

Of course, before we saw the data we would not know about this pattern so it could be argued that we should use a more general model in which we allow $q_{1,1} > 0$. In this case we would have

$$\begin{pmatrix} q_{1,1} & q_{1,2} \\ q_{2,1} & q_{2,2} \end{pmatrix} = \begin{pmatrix} 1-\pi_2 & \pi_1 \\ \pi_2 & 1-\pi_1 \end{pmatrix}. \tag{4.6}$$

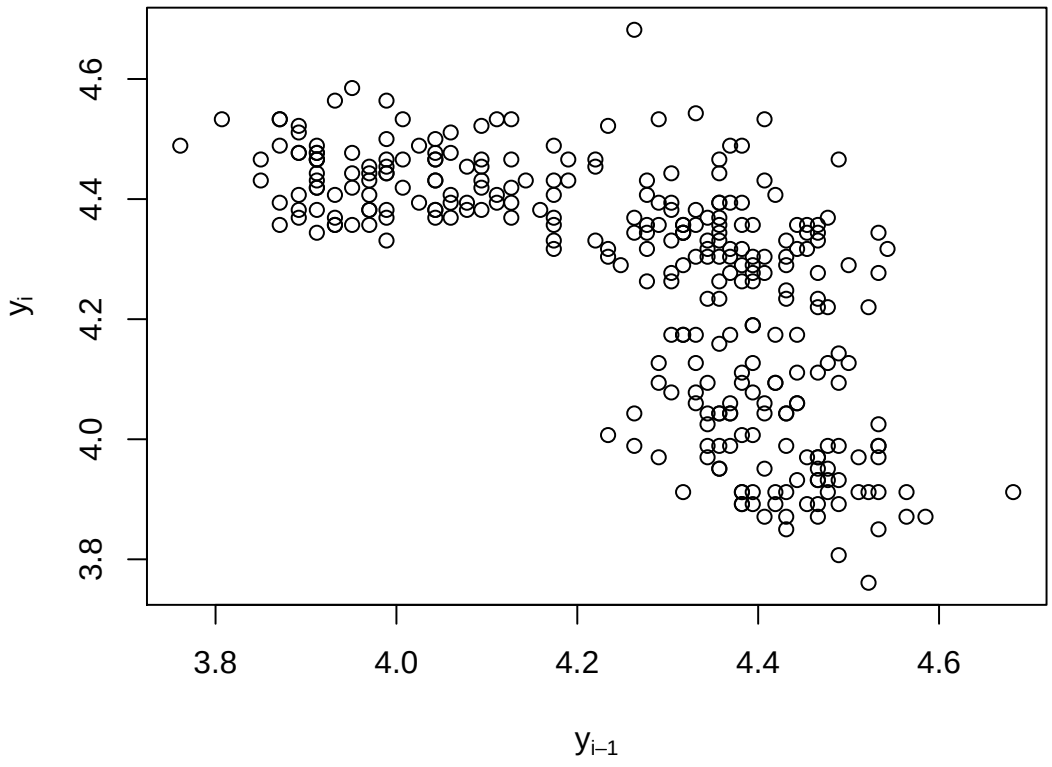

Figure 4.5: Logarithms of time intervals between eruptions of "Old Faithful". The i^{th} log interval y_i is plotted against the preceding log interval y_{i-1}.

This is an example of a *hidden Markov model* or HMM. In this case there are two *hidden states*. There are many different kinds of HMM and they have many applications, in such diverse areas as time series, DNA sequences and linguistics. In general, in a HMM, we have a sequence of (possibly vector) observations $\ldots y_{i-1}, y_i, y_{i+1} \ldots$ where the distribution of y_i depends on the value of an unobserved (i.e. *latent*) (possibly vector) variable x_i and the sequence $\ldots x_{i-1}, x_i, x_{i+1}, \ldots$ forms a Markov chain. There may, of course, be more than two hidden states.

Figure 4.6 shows a DAG for a typical HMM. There will typically also be unknown parameters on which the distributions depend but these have been omitted. Notice that (given the model parameters) the observations Y only depend on each other through the latent variables X. Figure 4.7 shows a DAG with the addition of the unknown parameters $\underline{\mu} = (\mu_1, \mu_2)^T$, $\underline{\tau} = (\tau_1, \tau_2)^T$ and π, for a case where the conditional distribution of Y_i when $X_i = c_i$ is $N(\mu_{c_i}, \tau_{c_i}^{-1})$ for $c_i = 1, 2$.

4.3.2 Two-state hidden Markov model

In the Old Faithful model, the latent variable X_i is the component membership c_i and there are two components. This is an example of a two-state HMM.

The transition matrix (4.6) defines the conditional distribution of c_i given c_{i-1}. To complete the model specification we have to give a distribution to the initial state c_1 (or to c_0, the state immediately before the start of the data). Very often we regard the process as stationary. That is the properties are not changing over time. In this case the initial state should have the stationary distribution of the Markov chain which can be found by solving

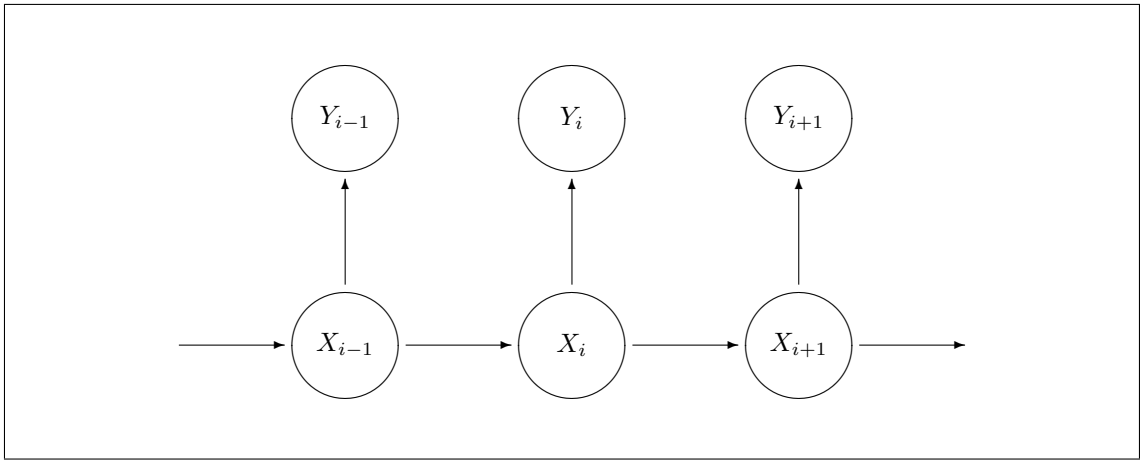

Figure 4.6: Directed acyclic graph for hidden Markov model

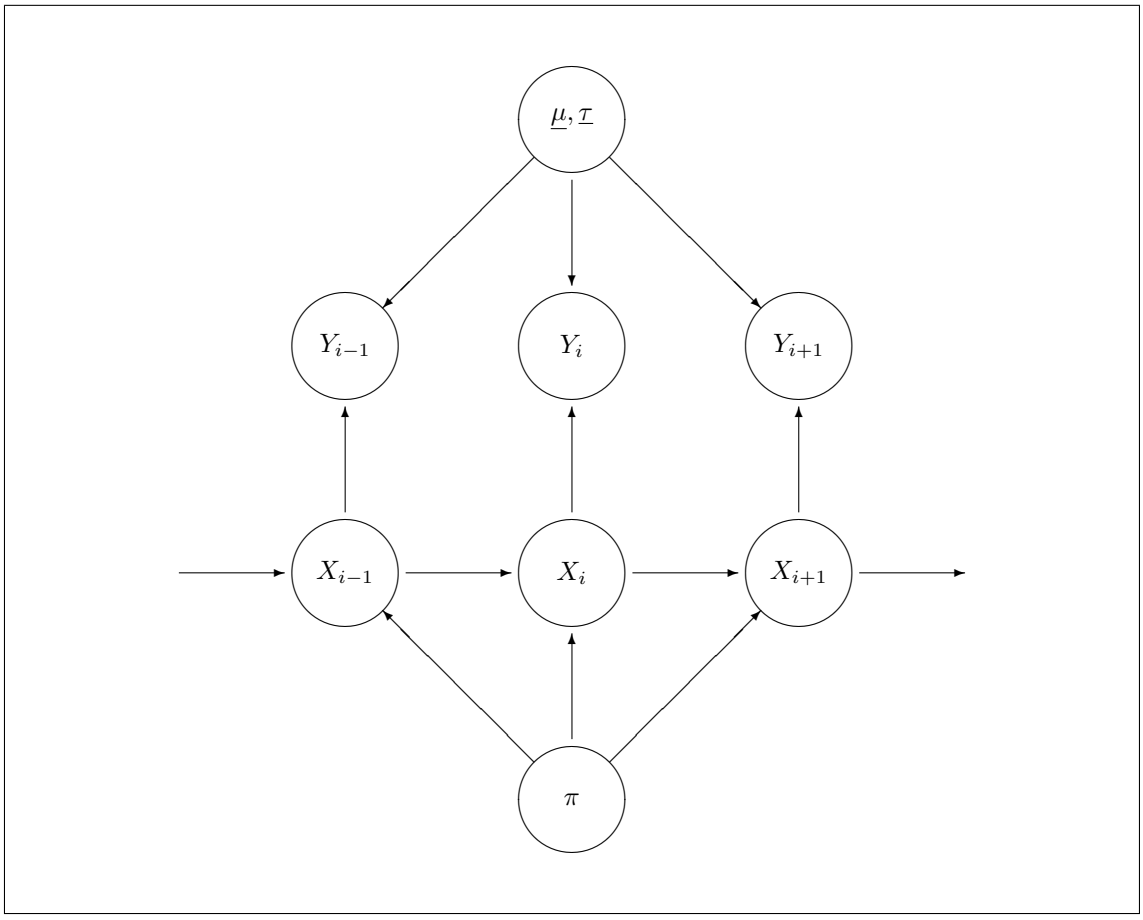

Figure 4.7: Directed acyclic graph for hidden Markov model showing unknown parameters

$$\begin{pmatrix} 1-\pi_2 & \pi_1 \\ \pi_2 & 1-\pi_1 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}$$

for P_1 and P_2 with the constraint that $P_1 + P_2 = 1$. The solution is

$$P_1 = \Pr(c_1 = 1) = \frac{\pi_1}{\pi_1 + \pi_2}, \quad P_2 = \Pr(c_1 = 2) = \frac{\pi_2}{\pi_1 + \pi_2}. \tag{4.7}$$

In the past we had two problems with this when using BRugs software. These may or may not still apply when using `rjags`.

1. BUGS software can not (or could not) handle the resulting likelihood with the complication of (4.7). We can, of course, write a Gibbs or Metropolis-Hastings algorithm of our own in R or some other programming language. However we could also use a trick to make BUGS work and which would give the correct result to a good approximation. The trick is to add the states $c_{-s}, \ldots, c_0$ as auxiliary variables for some reasonably large s (e.g. 30). We then give c_{-s} a convenient distribution, e.g. $\Pr(c_{-s} = 1) = \Pr(c_{-s} = 2) = 0.5$, or even just fix its value. The choice of this distribution or value has little effect on the posterior distribution. (This can be checked numerically). Figure 4.8 shows a BUGS model specification for the "Old Faithful" model, using (4.5) and normal distributions for the log interval times.

2. The BUGS model shown in Figure 4.8 worked satisfactorily in previous years. However, this year, it causes R to crash. This seems to be associated with a change in the version of R. Perhaps a change needs to be made to the BRugs package and this has not been made.

Because of these problems, especially Point 2, we will not attempt to use BRugs for hidden Markov models this year but, instead use specially-written R functions to demonstrate the use of MCMC with these models. In many cases we can sample from most of the fcds straightforwardly. The one exception is sampling values for π_1 and π_2 because of the term for the initial state. However we can use a Metropolis-Hastings sampler for this and thus have a Metropolis-within-Gibbs scheme.

For sampling π_1 and π_2 we can proceed as follows.

Suppose that the current value of the state at time 1 is c_1. Let

$$P(c_1, \pi_1, \pi_2) = \frac{\pi_{c_1}}{\pi_1 + \pi_2}.$$

Let the current numbers of state transitions, according to the currently allocated states, be $n_{1,1}, n_{1,2}, n_{2,1}, n_{2,2}$. That is, according to the allocation of observations to states at this iteration, we have $n_{j,k}$ transitions from state k to state j. Let $L^\star$ be the "likelihood" based just on these transitions. Then

$$L^\star = (1-\pi_2)^{n_{1,1}} \pi_2^{n_{2,1}} \pi_1^{n_{1,2}} (1-\pi_1)^{n_{2,2}}.$$

Suppose, for example, that we have independent beta prior distributions for π_1 and π_2, so that the joint prior density for π_1 and π_2 is

$$g_1^{(0)}(\pi_1) g_2^{(0)}(\pi_2) \propto \pi_1^{a_1-1} (1-\pi_1)^{b_1-1} \pi_2^{a_2-1} (1-\pi_2)^{b_2-1}.$$

Then, based just on $L^\star$, the joint "posterior" density is $g_1^{(1)}(\pi_1) g_2^{(1)}(\pi_2)$ where $g_1^{(1)}(\pi_1)$ is the density of a Beta$(a_1 + n_{1,2},\ b_1 + n_{2,2})$ distribution and $g_2^{(1)}(\pi_2)$ is the density of a Beta$(a_2 + n_{2,1},\ b_2 + n_{1,1})$ distribution. As a proposal, we can sample values $\pi_{1,\text{prop}}$ and $\pi_{2,\text{prop}}$ for π_1 and π_2 from this joint "posterior". That is we take independent samples from the two beta distributions.

However the fcd has density $k(n_{1,1}, n_{1,2}, n_{2,1}, n_{2,2}) P(c_1, \pi_1, \pi_2) g_1^{(1)}(\pi_1) g_2^{(1)}(\pi_2)$ where the constant $k(n_{1,1}, n_{1,2}, n_{2,1}, n_{2,2})$ does not depend on π_1 or π_2. Therefore, if $\pi_{1,\text{old}}$ and $\pi_{2,\text{old}}$ are the current values, from the preceding iteration, the acceptance ratio is

```
model

{p0[1]<-0.5
 p0[2]<-0.5
 cc[1]~dcat(p0[])

 for (i in 2:30)
   {cc[i]~dcat(q[,cc[i-1]])   # This is the "burn-in"section.
    }

 c[1]~dcat(q[,cc[30]])     # This is for the initial state.

 for (i in 2:n)
   {c[i]~dcat(q[,c[i-1]])
    }

 for (i in 1:n)
   {y[i]~dnorm(mu[c[i]],tau[c[i]])
    }

 for (j in 1:2)
   {tau[j]~dgamma(4,0.04)
    }

 mumean[1]<-mu0-0.2
 mumean[2]<-mu0+0.2
 for (j in 1:2)
     {mudash[j]~dnorm(mumean[j],3.3)
      }
 mu[1:2]<-sort(mudash)  # This imposes the order constraint.

 mu0~dnorm(4.0,p.mu)
 p.mu<-1/0.3

 q[1,2]<-pi
 pi~dbeta(1,1)
 q[2,2]<-1-q[1,2]
 q[1,1]<-0.0
 q[2,1]<-1-q[1,1]

 }
```

Figure 4.8: BUGS model specification for “Old Faithful” normal hidden Markov model.

$$\begin{aligned} A &= \frac{k(n_{1,1},n_{1,2},n_{2,1},n_{2,2})P(c_1,\pi_{1,\text{prop}},\pi_{2,\text{prop}})g_1^{(1)}(\pi_{1,\text{prop}})g_2^{(1)}(\pi_{2,\text{prop}})}{k(n_{1,1},n_{1,2},n_{2,1},n_{2,2})P(c_1,\pi_{1,\text{old}},\pi_{2,\text{old}})g_1^{(1)}(\pi_{1,\text{old}})g_2^{(1)}(\pi_{2,\text{old}})} \\ &\quad \times \frac{g_1^{(1)}(\pi_{1,\text{old}})g_2^{(1)}(\pi_{2,\text{old}})}{g_1^{(1)}(\pi_{1,\text{prop}})g_2^{(1)}(\pi_{2,\text{prop}})} \\ &= \frac{P(c_1,\pi_{1,\text{prop}},\pi_{2,\text{prop}})}{P(c_1,\pi_{1,\text{old}},\pi_{2,\text{old}})}. \end{aligned}$$

When sampling the component memberships c_i, the fcd depends on the transition probability from the preceding state, the transition probability to the succeeding state and the conditional density of the observation. So, for example, suppose that the conditional distributions are normal with means μ_1 and μ_2 and precisions τ_1 and τ_2 and the observation is y_i. Then write the conditional densities as $f(y_i;\mu_1,\tau_1)$ and $f(y_i;\mu_2,\tau_2)$. Given that the preceding state is c_{i-1} and the succeeding state is c_{i+1} and the transition matrix is as given in (4.6), then the "prior probability" that $c_i = j$ is proportional to

$$\Pr(c_i = j \mid c_{i-1}) \Pr(c_{i+1} \mid c_i = j) = q_{j,c_{i-1}} q_{c_{i+1},j}.$$

Multiplying "prior" by "likelihood" we find that the fcd probability that $c_i = j$ is proportional to $q_{j,c_{i-1}} q_{c_{i+1},j} f(y_i;\mu_j,\tau_j)$. Therefore the fcd probability that $c_i = j$ is

$$\frac{q_{j,c_{i-1}} q_{c_{i+1},j} f(y_i;\mu_j,\tau_j)}{\sum_{c=1}^{2} q_{c,c_{i-1}} q_{c_{i+1},c} f(y_i;\mu_c,\tau_c)}.$$

Figures 4.9 and 4.10 show a R function for a two-state HMM, as developed here, with normal conditional distributions. The conjugate normal-gamma form is used for the prior for each normal distribution. Note that the R command `table` produces the transpose of the table of counts used in these notes. Therefore, in the R function `hmmnorm`, the variables `ns[2,1]` and `ns[1,2]` correspond to $n_{1,2}$ and $n_{2,1}$ respectively.

4.3.3 Forward-backward algorithm

Mixing can be poor when using a Gibbs sampler with a HMM if we sample the hidden state at each time point separately. This is because there can be strong correlation in the posterior distribution between the states at neighbouring time points. We can overcome this problem by sampling the whole collection of hidden states as a block. This can be done using a procedure called the *forward-backward algorithm*. We do not have time to cover this in this course. See, for example, Scott (2002).

4.4 Practical 4

4.4.1 Simulated normal mixture data

Mixture models can sometimes be tricky to fit so we will start with an artificial example which is deliberately made so that it will work well.

The data `mixturedata.txt` and the BUGS model file `mixturenormbug.txt` can both be obtained from the web page. To read the data in R, type:

```
source("mixturedata.txt")
```

1. It is often necessary to help the software by providing initial values for the Gibbs sampler. In the case of mixture models it is also particularly important to check convergence. One way to do this is to run the sampler more than once, starting with different initial values. Therefore we create two sets of initial values.

   ```
   mixturenorminits<-list(list(mu=c(1,7),pi=0.3),list(mu=c(1,7),pi=0.7))
   ```

```
hmmnorm<-function(niter,prior,y)
{n<-length(y)
 cv<-ifelse(y<mean(y),1,2)                    # Initialise component memberships.
 m<-  matrix(nrow=2,ncol=2)
 mu<- matrix(nrow=niter,ncol=2)
 tau<-matrix(nrow=niter,ncol=2)
 pi<- matrix(nrow=niter,ncol=2)
 proportion<-numeric(niter)
 piprop<-numeric(2)
 piold <-c(0.5,0.5)
 for (iter in 1:niter)
     {ns<-table(cv[1:(n-1)],cv[2:n])                                    # Count the transitions.
      piprop[1]<-rbeta(1,prior$a[1]+ns[2,1],prior$b[1]+ns[2,2])         # Proposal for pi_1.
      piprop[2]<-rbeta(1,prior$a[2]+ns[1,2],prior$b[2]+ns[1,1])         # Proposal for pi_2.
      Pprop<-piprop[cv[1]]/sum(piprop)                                  # Stationary probability.
      Pold <-piold[cv[1]]/sum(piold)                                    # Stationary probability.
      A<-min(1,Pprop/Pold)                                              # Acceptance probability.
      U<-runif(1)
      if (U<A)                                                          # M-H sampling for pi.
         {pi[iter,]<-piprop
          piold<-piprop
          }
      else
         {pi[iter,]<-piold
          }
      m[1,1]<-1-pi[iter,2]                                              # Transition matrix.
      m[1,2]<-pi[iter,1]
      m[2,1]<-pi[iter,2]
      m[2,2]<-1-pi[iter,1]
      for (comp in 1:2)                                                 # Sample other parameters.
          {nc<-sum(cv==comp)
           if (nc>0)
              {ycomp<-y[cv==comp]
               ybar<-mean(ycomp)
               s2n<-(sum(ycomp*ycomp)-nc*ybar*ybar)/nc
               ycd<-ycomp-prior$m[comp]
               r2<-sum(ycd*ycd)/nc
               k1<-prior$c[comp]+nc
               d1<-prior$d[comp]+nc
               m1<-(prior$c[comp]*prior$m[comp]+nc*ybar)/k1
               vd<-(prior$c[comp]*r2+nc*s2n)/k1
               v1<-(prior$d[comp]*prior$v[comp]+nc*vd)/d1
               tau[iter,comp]<-rgamma(1,(d1/2),(d1*v1/2))
               sd<-sqrt(1/(k1*tau[iter,comp]))
               mu[iter,comp]<-rnorm(1,m1,sd)
               }
            else
              {tau[iter,comp]<-rgamma(1,(prior$d[comp]/2),(prior$d[comp]*prior$v[comp]/2))
               sd<-sqrt(1/(prior$c[comp]*tau[iter,comp]))
               mu[iter,comp]<-rnorm(1,prior$m[comp],sd)
               }
          }
```

Figure 4.9: R function for a two-state HMM with normal conditional distributions (Part 1).

```
      P<-pi[iter,]/sum(pi[iter,])                                    # Stationary probabilities.
      pc<-P*m[cv[2],]                                                # "Prior probs" for cv_1.
      sd<-numeric(2)
      for (comp in 1:2)
          {sd[comp]<-sqrt(1/tau[iter,comp])
           pc[comp]<-pc[comp]*dnorm(y[1],mu[iter,comp],sd[comp])     # "Prior times likelihood".
           }
      pc<-pc/sum(pc)                                                 # Normalise.
      cv[1]<-2-rbinom(1,1,pc[1])                                     # Sample cv_1.
      for (t in 2:(n-1))
          {pc<-m[,cv[t-1]]*m[cv[t+1],]                               # "Prior probs" for cv_t.
           for (comp in 1:2)
               {pc[comp]<-pc[comp]*dnorm(y[t],mu[iter,comp],sd[comp])   # "Prior times likelihood".
                }
           pc<-pc/sum(pc)                                            # Normalise.
           cv[t]<-2-rbinom(1,1,pc[1])                                # Sample cv_t.
           }
       pc<-m[,cv[n-1]]                                               # "Prior probs" for cv_n.
       for (comp in 1:2)
           {pc[comp]<-pc[comp]*dnorm(y[n],mu[iter,comp],sd[comp])    # "Prior times likelihood".
            }
       pc<-pc/sum(pc)                                                # Normalise.
       cv[n]<-2-rbinom(1,1,pc[1])                                    # Sample cv_n.
       proportion[iter]<-sum(cv==1)/n                                # Proportion in component 1.
       }
out<-list(mu=mu,tau=tau,pi=pi,proportion=proportion)
out
 }
```

Figure 4.10: R function for a two-state HMM with normal conditional distributions (Part 2).

So, we will start with very different probabilities of an observation being in component 1.

2. Build the `rjags` model and check the convergence of `pi`, the probability for component 1. We will set the two different intial values in two separate chains and run them without burn-in periods.

```
mixturejags<-jags.model("mixturenormbug.txt",data=mixturedata,inits=mixturenorminits,n.chains=2)
mixturesamples<-coda.samples(mixturejags,c('pi'),1000)
par(ask=TRUE)
traceplot(mixturesamples)
```

Look at the resulting graph. You should see that "convergence" has been quite quick.

3. Compute the posterior distribution. This time we will use a burn-in.

```
mixturejags<-jags.model("mixturenormbug.txt",data=mixturedata,inits=mixturenorminits,n.chains=2)
update(mixturejags,1000)
mixturesample<-coda.samples(mixturejags,c('pi'),2000)
```

4. Look at the results. For example:

```
summary(mixturesamples)
densplot(mixturesamples)
```

4.4.2 Old Faithful log-normal mixture

We will fit a two-component normal mixture to the logs of the intervals between eruptions of "Old Faithful."

The data `geyserlogdata.txt` and the BUGS model file `faithnormbug.txt` can both be obtained from the web page. To read the data, type:

```
source("geyserlogdata.txt")
```

Create two different sets of initial value files.

```
geyserloginits<-list(list(mu=c(4.0,4.4),pi=0.3),list(mu=c(4.0,4.4),pi=0.7))
```

So, we will start with very different probabilities of an observation being in component 1.

1. Check convergence of `pi`, the probability for component 1. We will set the two different intial values in two separate chains and run them without burn-in periods.

```
mixturelogjags<-jags.model("faithnormbug.txt",data=geyserlogdata,inits=geyserloginits,n.chains=2)
mixturelogsamples<-coda.samples(mixturelogjags,c('pi'),1000)
par(ask=TRUE)
traceplot(mixturesamples)
```

Look at the resulting graph. Has "convergence" been quick?

2. Compute the posterior distribution. This time we will use a burn-in.

```
mixturelogjags<-jags.model("faithnormbug.txt",data=geyserdata,inits=geyserloginits,n.chains=2)
update(mixturelogjags,1000)
mixturelogsamples<-coda.samples(mixturejags,c('pi','mu','tau'),2000)
```

3. Look at the results. For example:

```
summary(mixturelogsamples)
par(ask=TRUE)
densplot(mixturelogsamples)
```

4.4.3 Old Faithful gamma mixture

We will fit a two-component gamma mixture to the intervals between eruptions of "Old Faithful."

The data `geyserdata.txt` and the BUGS model file `faithgammabug.txt` can both be obtained from the web page.

We can use the same initial value files as we used for the normal mixture.

1. Check convergence of `pi`, the probability for component 1. We will set the two different intial values in two separate chains and run them without burn-in periods.

```
modelCheck("faithgammabug.txt")
modelData("geyserdata.txt")
modelCompile(2)
modelInits("faithnorminits1.txt")
modelInits("faithnorminits2.txt")
modelGenInits()
samplesSet("pi")
modelUpdate(1000)
samplesHistory("pi")
```

   Look at the resulting graph. Has "convergence" been quick?

2. Compute the posterior distribution. This time we will use a burn-in.

```
modelCheck("faithgammabug.txt")
modelData("geyserdata.txt")
modelCompile(2)
modelInits("faithnorminits1.txt")
modelInits("faithnorminits2.txt")
modelGenInits()
modelUpdate(1000)
samplesSet(c("pi","alpha","beta"))
modelUpdate(3000)
```

3. Look at the results. For example:

```
samplesStats(c("pi","alpha","beta"))
samplesDensity("pi")
samplesDensity("alpha")
samplesDensity("beta")
```

4. The marginal posterior distributions for β_1 and β_2 are quite similar. Perhaps values of β_1 and β_2 which are close to each other would represent the data well. Let us look at the posterior distribution of β_1/β_2.

```
beta1<-samplesSample("beta[1]")
beta2<-samplesSample("beta[2]")
betaratio<-beta1/beta2
plot(density(betaratio))
```

4.4.4 Road traffic headways (independent)

We will fit an exponential/gamma mixture model to some road traffic headway data. Two files of data are avilable on the web page. They are:

```
dd01data.txt
feb28peledata.txt
```

The first was collected by my research student, Ged Cowburn. I collected the second. You can use either one. They have slightly different characteristics.

The BUGS model file is also available on the Web page as `headway0bug.txt`.

We will need some initial value files. Create two files as follows.

`headwayinits1.txt`

containing

```
list(aa=2, bb=0.5, pi=0.1)
```

and

`headwayinits2.txt`

containing

```
list(aa=2, bb=0.5, pi=0.7)
```

1. Check convergence of `pi` , the probability for component 1. We will set the two different intial values in two separate chains and run them without burn-in periods. I will use `dd01data.txt` but you can use `feb28peledata.txt` if you wish.

   ```
   modelCheck("headway0bug.txt")
   modelData("dd01data.txt")
   modelCompile(2)
   modelInits("headwayinits1.txt")
   modelInits("headwayinits2.txt")
   modelGenInits()
   samplesSet("pi")
   modelUpdate(1000)
   samplesHistory("pi")
   ```

   Look at the resulting graph. You will probably see that the samplers have "converged" but that "mixing" is not very good. Therefore we need to take a large number of samples.

2. Compute the posterior distribution. This time we will use a burn-in.

   ```
   modelCheck("headway0bug.txt")
   modelData("dd01data.txt")
   modelCompile(2)
   modelInits("headwayinits1.txt")
   modelInits("headwayinits2.txt")
   modelGenInits()
   modelUpdate(2000)
   samplesSet(c("pi","alpha","beta"))
   modelUpdate(10000)
   ```

3. Look at the results. For example:

   ```
   samplesStats(c("pi","alpha","beta"))
   samplesDensity("pi")
   samplesDensity("alpha")
   samplesDensity("beta")
   ```

4.4.5 Old Faithful (log-normal hidden Markov model)

Try fitting the log-normal hidden Markov model to the Old Faithful data. The R function shown in Figures 4.9 and 4.10 is available on the Web page as `hmmR.txt`. It seems to work well despite the fact that I have not used any defence against label-switching, other than giving the two components different prior means and initialising the component memberships to favour the correct allocation. The logs of the intervals, in a suitable form, are in a file `geyserlogdatab.txt` on the Web page.

1. Load the data:

```
y<-scan("geyserlogdatab.txt")
```

2. Set up the prior:

```
a<-c(1,1)
b<-c(1,1)
m<-c(3.5,4.5)
d<-c(8,8)
v<-c(0.01,0.01)
c<-c(0.01,0.01)
prior<-list(a=a,b=b,m=m,d=d,v=v,c=c)
```

3. Install the function:

```
source("hmmR.txt")
```

4. Try, for example, 1000 iterations:

```
test<-hmmnorm(1000,prior,y)
```

5. Have a look at the results. For example:

```
mu<-test$mu
Iteration<-1:1000
plot(Iteration,mu[,1],type="l",col=2,ylim=c(3.5,5))
lines(Iteration,mu[,2],col=3)
pi<-test$pi
plot(Iteration,pi[,1],type="l",col=2,ylim=c(0.0,1.0))
lines(Iteration,pi[,2],col=3)
plot(density(mu[,1]))
```

   In particular, notice that, as expected π_2 turns out to be close to 1.

6. Try anything else you like.

Note that the function gives a single chain. If you want to try multiple chains, you have to run the function multiple times.

4.4.6 Headways (hidden Markov model)

Cowburn and Farrow (2007) discussed fitting hidden Markov models to series of road-vehicle headways. The two component distributions were as in section 4.2.2. The transition matrix was as in (4.6).

This model is complicated by the fact that both parameters are unknown in one of the conditional gamma distributions and sampling the fcd for the "shape" or "index" parameter (*ie* the first parameter) is not straightforward. To avoid this complication we will fix its value at 4.0.

You can download a suitable R function from the file `hmmheadwayR.txt` on the Web page. I have marked with `######` the places where changes have been made from `hmmR`.The first set of headway data is available in the file `dd01datab.txt` on the Web page. (You could also easily edit the other set to make it usable this way if you so wished). The function seems to work well even though, again, I have not really defended against label switching.

1. Load the data:

```
t<-scan("dd01datab.txt")
```

2. Set up the prior:

```
a<-c(1,1)
b<-c(2,2)
abeta<-c(2,1)
bbeta<-c(8,0.5)
priorh<-list(a=a,b=b,abeta=abeta,bbeta=bbeta)
```

3. Install the function:

```
source("hmmheadwayR.txt")
```

4. Try, for example, 1000 iterations:

```
test<-hmmheadway(1000,priorh,t)
```

5. Have a look at the results. For example:

```
mean<-test$mean
Iteration<-1:1000
plot(Iteration,mean[,1],type="l",col=2,ylim=c(0,20),ylab="Mean headway")
lines(Iteration,mean[,2],col=3)
lrr<-log(test$pi[,2]/(1-test$pi[,1]))
plot(density(lrr))
```

   I expect that you will see that a short burn-in might help. You can easily delete the first few iterations from the output.

   The quantity `lrr` is the log relative risk for being in component 2 next time comparing being in component 1 now with being in component 2 now. As you can see, there is little evidence that this differs much from zero. Thus there is little evidence that the component memberships are not independent and that we need a hidden Markov model at all. At least, this is what is suggested by this model!

6. Try anything else you like.

Chapter 5

Random Effects and Hierarchical Models

5.1 Random Effects

5.1.1 Fixed and random effects

Consider Example 2 of Lecture 1.3.1. The data gave the gains in weight of rats fed on four different diets. The diets differed in terms of the amount of protein ("low" or "high") and the source of the protein ("beef" or "cereal"). The population mean weight gains with each diet are considered to be parameters of the model. If we were to observe very large numbers of rats with each diet then we would gain very precise information about the values of these parameters. In the limit, we would know the values exactly. The differences in population mean weight gains between the diets are regarded as *fixed* but *unknown*.

We can write the four means in the form

$$\begin{aligned}
\mu_1 &= \mu - \beta_a - \beta_s + \gamma, \\
\mu_2 &= \mu + \beta_a - \beta_s - \gamma, \\
\mu_3 &= \mu - \beta_a + \beta_s - \gamma, \\
\mu_4 &= \mu + \beta_a + \beta_s + \gamma.
\end{aligned}$$

Then $\beta_a,\ \beta_s, \gamma$ are all regarded as *fixed effects*.

Now consider another example. This example comes from Davies and Goldsmith (1972). The experiment concerned testing the strength of Portland cement. The cement was divided into small samples. Each sample was then mixed with water and worked. This process is called "gauging." Each sample was then cast into a cube and allowed to set. The samples were then tested for strength. This is known as "breaking."

Three different people did the gauging and three different people did the breaking. There are thus nine combinations of gauger and breaker. In each combination there were four cubes. The data, in pounds per square inch, are given in Table 5.1.

Let $y_{i,j,k}$ be the i^{th} observation made with Gauger j and breaker k, a realisation of the random variable $Y_{i,j,k}$. Then we might write

$$Y_{i,j,k} \mid \mu_{j,k}, \tau_\varepsilon \sim N(\mu_{j,k},\ \tau_\varepsilon^{-1})$$

where

$$\mu_{j,k} = \mu + \alpha_j + \beta_k + \gamma_{j,k}.$$

Here, just as in the rats example, we have the main effects of two factors and an interaction effect. The gauger effects are $\alpha_1, \alpha_2, \alpha_3$, the breaker effects are $\beta_1, \beta_2, \beta_3$ and the interaction effects are $\gamma_{1,1}, \ldots, \gamma_{3,3}$. However these are not regarded as *fixed* effects. Instead they are regarded as *random effects*. This is because we are not just interested in the effects of *these* gaugers and

	Breaker 1		Breaker 2		Breaker 3	
Gauger 1	5280	5520	4340	4400	4160	5180
	4760	5800	5020	6200	5320	4600
Gauger 2	4420	5280	5340	4880	4180	4800
	5580	4900	4960	6200	4600	4480
Gauger 3	5360	6160	5720	4760	4460	4930
	5680	5500	5620	5560	4680	5600

Table 5.1: Breaking strengths (pounds per square inch) of cement samples.

breakers but in how much variation there is between gaugers generally and between breakers generally. We regard these gaugers as a sample from the population of gaugers and these breakers as a sample from the population of breakers.

We do not constrain the effects to sum to zero, or fix one of them to be zero. Instead we regard them as samples from a distribution with zero mean. The mean is zero because we include the parameter μ which absorbs any nonzero mean.

In this example, and this is typical, we say that, given the precision parameters $\tau_\alpha,\ \tau_\beta,\ \tau_\gamma$,

$$\begin{aligned} \alpha_j &\sim N(0,\ \tau_\alpha^{-1}) \quad (j = 1, \ldots, 3) \\ \beta_k &\sim N(0,\ \tau_\beta^{-1}) \quad (k = 1, \ldots, 3) \\ \gamma_{j,k} &\sim N(0,\ \tau_\gamma^{-1}) \quad (j = 1, \ldots, 3,\ k = 1, \ldots, 3) \end{aligned}$$

We then give prior distributions to the model parameters $\mu,\ \tau_\alpha, \tau_\beta,\ \tau_\gamma,\ \tau_\varepsilon$. Typically

$$\begin{aligned} \mu &\sim N(m_0,\ v_0), \\ \tau_\alpha &\sim \text{Ga}(a_\alpha,\ b_\alpha), \\ \tau_\beta &\sim \text{Ga}(a_\beta,\ b_\beta), \\ \tau_\gamma &\sim \text{Ga}(a_\gamma,\ b_\gamma), \\ \tau_\varepsilon &\sim \text{Ga}(a_\varepsilon,\ b_\varepsilon). \end{aligned}$$

Inference involves using data to learn about these population parameters. The variances $\tau_\alpha^{-1},\ \tau_\beta^{-1},\ \tau_\gamma^{-1},\ \tau_\varepsilon^{-1}$ are known as *variance components*.

Notice two difference between this random effects model and the fixed effects model which we used for the rats example.

1. We suppose that we might observe a new gauger or a new breaker in the future. No matter how many observations we make with the gaugers and breakers in our sample, we will never be able to predict exactly the mean for a new gauger-breaker combination which have not yet observed because this will involve new realisations from the random effects distributions.

2. Suppose that, instead of giving $\tau_\alpha,\ \tau_\beta,\ \tau_\gamma$ prior distributions, we simply chose values for them. Then the model would be very similar to a fixed effects model. The only differences would be point 1, which refers to how we interpret the results in terms of future observations, and the fact that we do not constrain the effects to sum to zero. This latter point would mean that the individual model effects would not be identifiable but the nine means for combinations of gauger and breaker would still be identifiable. However, in fact, we do not choose values for these precisions (i.e. for the variance components) but regard them as unknown and learn about them from the data. This means that we use the data to tell us how similar we can expect future gauger-breaker combinations to be to those which we have already seen.

5.1.2 Evaluation of posterior distribution

Given a model such as the cement-testing example, we can easily use MCMC with data augmentation to sample from the posterior distribution. We regard the random effects as auxiliary variables. I will illustrate the method in terms of the cement example. The auxiliary data are $\alpha_1,\ldots,\alpha_3,\ \beta_1,\ldots,\beta_3,\gamma_{1,1},\ldots,\gamma_{3,3}$.

A possible MCMC scheme is as follows. Sketching a DAG might help to see how this works.

1 Sample τ_ε : Given values for the fixed effect μ and the random effects we have

$$Y_{i,j,k}-\mu-\alpha_j-\beta_k-\gamma_{j,k}\sim N(0,\ \tau_\varepsilon^{-1})$$

With a gamma prior for τ_ε we get a gamma fcd for τ_ε and it is easy to sample from this.

2 Sample μ : Given values for the error precision τ_ε and the random effects we have

$$Y_{i,j,k}-\alpha_j-\beta_k-\gamma_{j,k}\sim N(\mu,\ \tau_\varepsilon^{-1})$$

With a normal prior for μ we get a normal fcd for μ and it is easy to sample from this.

3 Sample τ_α : Given τ_α we have $\alpha_j\sim N(0,\ \tau_\alpha^{-1})$. So, given values for $\alpha_1,\ldots,\alpha_3$ and a gamma prior for τ_α, the fcd for τ_α is a gamma distribution and it is easy to sample from this.

4 Sample τ_β : Given τ_β we have $\beta_k\sim N(0,\ \tau_\beta^{-1})$. So, given values for $\beta_1,\ldots,\beta_3$ and a gamma prior for τ_β, the fcd for τ_β is a gamma distribution and it is easy to sample from this.

5 Sample τ_γ : Given τ_γ we have $\gamma_{j,k}\sim N(0,\ \tau_\gamma^{-1})$. So, given values for $\gamma_{1,1},\ldots,\gamma_{3,3}$ and a gamma prior for τ_γ, the fcd for τ_γ is a gamma distribution and it is easy to sample from this.

6 Sample $\alpha_1,\ldots,\alpha_3$: Given values for the fixed effect μ, for τ_α and for the other random effects we have

$$Y_{i,j,k}-\mu-\beta_k-\gamma_{j,k}\sim N(\alpha_j,\ \tau_\varepsilon^{-1})$$

The "prior" for α_j here is the conditional distribution of α_j given τ_α which is $\alpha_j\mid\tau_\alpha\sim N(0,\ \tau_\alpha^{-1})$. The resulting fcd is normal and it is easy to sample from this. The fcd for α_j just involves the data through $y_{1,j,1},\ldots,y_{4,j,3}$.

7 Sample $\beta_1,\ldots,\beta_3$: Given values for the fixed effect μ, for τ_β and for the other random effects we have

$$Y_{i,j,k}-\mu-\alpha_j-\gamma_{j,k}\sim N(\beta_k,\ \tau_\varepsilon^{-1})$$

The "prior" for β_k here is the conditional distribution of β_k given τ_β which is $\beta_k\mid\tau_\beta\sim N(0,\ \tau_\beta^{-1})$. The resulting fcd is normal and it is easy to sample from this. The fcd for β_k just involves the data through $y_{1,1,k},\ldots,y_{4,3,k}$.

8 Sample $\gamma_{1,1}, \ldots, \gamma_{3,3}$: Given values for the fixed effect μ, for τ_γ and for the other random effects we have

$$Y_{i,j,k} - \mu - \alpha_j - \beta_k \sim N(\gamma_{j,k},\ \tau_\varepsilon^{-1})$$

The "prior" for $\gamma_{j,k}$ here is the conditional distribution of $\gamma_{j,k}$ given τ_γ which is $\gamma_{j,k} \mid \tau_\gamma \sim N(0,\ \tau_\gamma^{-1})$. The resulting fcd is normal and it is easy to sample from this. The fcd for $\gamma_{j,k}$ just involves the data through $y_{1,j,k}, \ldots, y_{4,j,k}$.

Note that this is by no means the *only* way to evaluate the posterior distribution. In fact this algorithm may be subject to poor mixing. However it is simple to implement.

5.1.3 More general models

I have explained random effects models in terms of a simple example with two factors, each with three levels, and an interaction. Of course we could have much more complicated models with more factors and interactions. The principles remain the same though.

We could also have models which contain non-normal distributions We will see an example of this later.

5.1.4 Mixed models

We can also have models in which some effects are fixed and some random. For example, in testing two drugs for the control of high blood pressure, each patient might provide a number of blood pressure measurements while being treated with each of the drugs (e.g. in a crossover trial). We would normally regard the drug effects as fixed but the patient effects, and any patient-drug interaction, as random effects. At step 2 in the algorithm above we would sample all of the fixed effects.

A model containing both fixed and random effects is called a *mixed model*. A mixed model where the random effects distributions are normal and the error distribution is normal and the means are linear functions of the effects is a *linear mixed model*. We could also have, for example, a *generalised linear mixed model* in which the error distribution might be, for example, Poisson or binomial, the means are related to linear predictors by a link function and the linear predictors are linear functions of fixed and random effects.

5.2 Hierarchical Models

5.2.1 Hierarchical structures

We are going to look at models and priors where we have two or more "levels" of conditional distributions.

For example we might say that, given $A_{j,k}$ and σ_Y^2, we have

$$Y_{i,j,k} \mid A_{j,k}, \sigma_Y^2 \sim N(A_{j,k},\ \sigma_Y^2),$$

then, given B_k and σ_A^2, we have

$$A_{j,k} \mid B_k, \sigma_A^2 \sim N(B_k,\ \sigma_A^2),$$

then, given μ and σ_B^2, we have

$$B_k \mid \mu, \sigma_B^2 \sim N(\mu,\ \sigma_B^2).$$

We would then give priors to $\mu,\ \sigma_B^2,\ \sigma_A^2$ and σ_Y^2. Our prior for μ could be

$$\mu \sim N(m_0,\ v_0).$$

Notice that

- there are several *levels* in this structure and
- the structure is *nested* or *hierarchical*.

Here $Y_{i,j,k}$ is the i^{th} obervation within sub-group j of group k. Two observations within the same subgroup are more strongly correlated with each other than two observations within different subgroups. Two observations in different subgroups within the same group are more strongly correlated than two observations in different groups. The group means are themselves correlated in the prior.

It is easy to see that we could write (conditional on all of the variances)

$$Y_{i,j,k} = \mu + b_k + a_{j,k} + \varepsilon_{i,j,k}$$

where

$$\begin{aligned} \mu &\sim N(m_0,\ v_0), \\ b_k &\sim N(0,\ \sigma_B^2), \\ a_{j,k} &\sim N(0,\ \sigma_A^2), \\ \varepsilon_{i,j,k} &\sim N(0,\ \sigma_Y^2), \end{aligned}$$

all independently.

Therefore, in the example above:

- All of the observations $Y_{i,j,k}$ have prior mean m_0.
- The prior (predictive) variance for $Y_{i,j,k}$ is

$$V_Y = v_0 + \sigma_B^2 + \sigma_A^2 + \sigma_Y^2.$$

- The prior (predictive) covariance between $Y_{i,j,k}$ and $Y_{i',j,k}$, where $i' \neq i$, is

$$V_A = v_0 + \sigma_B^2 + \sigma_A^2.$$

- The prior (predictive) covariance between $Y_{i,j,k}$ and $Y_{i',j',k}$, where $j' \neq j$, is

$$V_B = v_0 + \sigma_B^2.$$

- The prior (predictive) covariance between $Y_{i,j,k}$ and $Y_{i',j',k'}$, where $k' \neq k$, is v_0.

When we looked at mixture models we said that there were two different reasons why we might use a mixture model, depending on whether or not we supposed that there really were subpopulations. The distinction between a "hierarchical prior" and a "multilevel model" or "hierarchical model" is of a similar nature.

In some cases we are really only interested in one level of unit, such as the sub-groups indexed j, k above, and other levels, e.g. the groups indexed k above, are introduced simply to give a covariance structure to the prior. In this case we would regard this as a "hierarchical prior."

In other cases the levels might have "physical" interpretations. For example, $Y_{i,j,k}$ could be the score obtained in a test by pupil i in school j of education authority k. Then the values of the education-authority effects b_k and the school effect $a_{j,k}$ might be of interest in themselves.

5.2.2 Hierarchical priors and "borrowing strength"

We have seen hierarchical priors already. We might make observations on members of a number of groups, e.g. weight gains of rats given different diets. So $Y_{i,j}$ is the i^{th} observation in Group j. Then, given μ_j and σ_Y^2, we have

$$Y_{i,j} \mid \mu_j, \sigma_Y^2 \sim N(\mu_j,\ \sigma_Y^2).$$

We need a prior for $\mu_1, \ldots, \mu_J$ but, if we are measuring the same thing in these groups, e.g. weight gain, then it seems reasonable that these means will be positively correlated in our prior. So, given μ_0 and σ_μ^2, we write

$$\mu_j \mid \mu_0, \sigma_\mu^2 \sim N(\mu_0,\ \sigma_\mu^2).$$

Then we give a prior to μ_0 with

$$\mu_0 \sim N(m_0,\ v_0).$$

Thus, in our prior, given σ_μ^2, each of μ_j and $\mu_{j'}$ has mean m_0 and variance $v_0 + \sigma_\mu^2$ but they also have covariance v_0 when $j \neq j'$.

Typically we would also give a prior to σ_Y^2. We might simply choose a value for σ_μ^2 but we might choose to give it a prior as well. Choosing to give σ_μ^2 a distribution has two effects.

- Because of the covariance structure, the posterior means of $\mu_1, \ldots, \mu_J$ will tend to be closer to their common overall mean than the sample means of the data are. This is a similar effect to the posterior mean being closer to the prior mean than the sample mean is when we have a single sample. This effect is called *shrinkage*. The degree of shrinkage depends, in part, on the relative sizes of the variances. If we choose the value of σ_μ^2 then we are (almost) choosing the degree of shrinkage. (The degree of shrinkage also depends on σ_Y^2 and we allow this to be unknown). If we allow σ_μ^2 to be unknown and give it a prior then we give the data more influence over the degree of shrinkage.

- If we expect to observe other related groups in the future then, learning about σ_μ^2 from the data allows us to change our minds about how close we expect these future group means to be to the means for groups which we have seen. We would have to believe that, in some sense, the future groups would be drawn from "the same population." (Usually this means that we would believe that groups were exchangeable).

Borrowing strength

The shrinkage effect noted above has an important benefit. Consider the following (very simplified) example.

We are interested in the rates of a disease in different areas of the country. In area j the population at risk is n_j. (In reality we would usually also take into account age groups etc.). Our model says that the number of cases in area j is Y_j which, conditional on a rate parameter θ_j, has a Poisson distribution

$$Y_j \mid \theta_j \sim \text{Po}(n_j\theta_j).$$

Now the mean of this distribution $n_j\theta_j$ might be a small number (e.g. 10) so that the standard deviation of the Poisson distribution is quite large compared to its mean. If we try to make inferences about the individual rates λ_j treating them independently then there is little information in the data about each. On the other hand, if we pool all of the data and assume that $\lambda_1 = \lambda_2 = \cdots = \lambda_J$, then we lose any possibility of detecting unusual rates in particular places. Instead we compromise and use a hierarchical prior. Given $a,\ b$ we give θ_j a gamma distribution

$$\theta_j \mid a, b \sim \text{Ga}(a,\ b).$$

we can then give a prior to $a,\ b$.

In this way the posterior distribution for θ_j uses information not only from Y_j but also from the observations in other areas. For example, the posterior means in cases with unusually large Y values will be "shrunk" somewhat towards the overall average. This is called "borrowing strength."

In *spatial statistics*, more complicated models are used in which the parameter in an area is more strongly correlated with the parameters in neighbouring areas.

5.2.3 Data augmentation and MCMC

Clearly, just as in Lecture 5.1 on random effects, hierarchical structures such as those discussed here give rise to a straightforward application of MCMC with data augmentation, regarding the different levels of random effects as auxiliary data. So, for example, in the first, normal, example above we could regard $\{A_{j,k}\}$ and $\{B_k\}$ as auxiliary data. For fixed values of these the likelihood is simple and sampling values for the parameters is simple. When the values of the parameters and one set of auxiliary variables is fixed, it is simple to sample values for the other set of auxiliary variables.

5.2.4 Multilevel models

As noted above, in some cases we are interested in the random effects themselves, rather than either just the population parameters ($\mu,\ \sigma_B^2,\ \sigma_A^2,\ \sigma_Y^2$) or just the first-level parameters ($\{A_{j,k}\}$).

5.3 Repeated Measures

5.3.1 Introduction

Among the types of problem where random effects are used are *repeated measures* models, where several observations are made on the same individual, and *longitudinal data*, where we are particularly interested in how repeated measurements taken on individuals change over time.

5.3.2 Example: Repeated measurements in two groups

A drug for lowering blood pressure is tested. A sample of patients with high blood pressure is divided randomly into two groups. Patients in Group 1 are given the drug. Patients in Group 2 are given a placebo. After a suitable period a sequence of five blood pressure measurements, at intervals, is made on each patient. (In this example we assume that there is no time trend).

This is really just a mixed-effects model. There is a fixed treatment effect and there are random patient effects.

Let $Y_{i,g,t}$ be the observation on patient i of group g at time t for $i = 1, \ldots, n_g,\ g = 1, 2,\ t = 1, \ldots 5$.

Model:

$$\begin{aligned} Y_{i,g,t} \mid P_{i,g}, \sigma_Y^2 &\sim N(P_{i,g},\ \sigma_Y^2) \\ P_{i,g} \mid \mu_g, \sigma_P^2 &\sim N(\mu_g,\ \sigma_P^2) \end{aligned}$$

Prior:

$$\begin{aligned} \mu_g \mid \mu_0, v_g &\sim N(\mu_0,\ v_g) \\ \mu_0 &\sim N(m_0,\ v_0) \\ \tau_Y = \sigma_Y^{-2} &\sim \text{Ga}(a_Y,\ b_Y) \\ \tau_P = \sigma_P^{-2} &\sim \text{Ga}(a_P,\ b_P) \end{aligned}$$

Notice that we are using a hierarchical prior for μ_1, μ_2. This gives them the same prior distribution. This might or might not seem to be a reasonable thing to do. Another possibility would be as follows.

$$\begin{aligned} \mu_1 &= \mu_0 + \delta \\ \mu_2 &= \mu_0 - \delta \\ \mu_0 &\sim N(m_0,\ v_0) \\ \delta &\sim N(d_0,\ \tilde{v}_g) \end{aligned}$$

```
model bloodpressure

{for (i in 1:N)
   {for (t in 1:5)
      {y[i,t]~dnorm(p_[i],tau.y)
       }
    p[i]~dnorm(mu[group[i]],tau.p)
    }

 for (g in 1:2)
   {mu[g]~dnorm(mu0,0.001)
    }

 mu0~dnorm(150,0.0005)
 tau.y~dgamma(2,100)
 tau.p~dgamma(2,200)
 }
```

Figure 5.1: BUGS code for blood pressure example.

Here the variance of $\mu_1 - \mu_2$ is $\text{var}(2\delta) = 4\tilde{v}_g$. In the first form of prior we had $\text{var}(\mu_1 - \mu_2) = 2v_g$. Hence, if we set $\tilde{v}_g = v_g/2$ we get the same variances and covariances. The introduction of d_0 allows us to have a nonzero prior mean for the treatment effect.

Figure 5.1 shows some suitable BUGS code. It is assumed that the data file contains six columns. The first column contains the number of the group to which the patient belongs. The remaining five columns contain the five blood pressure measurements, in order. (With BRugs it would be necessary to load the overall sample size `N` from another file). The first form of the prior is used.

5.3.3 Autocorrelation

In the example in 5.3.2 we have made no use of the time-ordering of the observations. The five observation on a particular patient are treated as exchangeable. We might believe that neighbouring observations are likely to be more strongly correlated than observations further apart. We could allow for this by allowing autocorrelation of the observations. This could be done, for example, using an autoregressive process or a moving average process. For illustration we will use a first-order moving average process.

The model as it stands can be written

$$Y_{i,g,t} = P_{i,g} + \varepsilon_{i,g,t}$$

where $\varepsilon_{i,g,t} \sim N(0,\ \sigma^2_Y)$.

Let us replace this with

$$Y_{i,g,t} = P_{i,g} + \varepsilon_{i,g,t} + \eta_{i,g,t} + \eta_{i,g,t+1}$$

where $\varepsilon_{i,g,t} \sim N(0, \sigma^2_\varepsilon)$ and $\eta_{i,g,t} \sim N(0,\ \sigma^2_\eta)$. The conditional variance of $Y_{i,g,t}$ is now $\sigma^2_\varepsilon + 2\sigma^2_\eta$ so we would want this variance to correspond to the old σ^2_Y. The conditional covariance between $Y_{i,g,t}$ and $Y_{i,g,t'}$ is now zero for $|t - t'| > 1$ but σ^2_η for $|t - t'| = 1$.

Figure 5.2 shows modified BUGS code. Note that we have to allow for an extra $\eta_{i,g,6}$.

5.3.4 Example: growth curves

Growth curves are a special kind of longitudinal-data problem. We are often interested in how, for example, individual children or young animals grow over time.

Here is a simple example taken from Gelfand *et al.* (1990). It can also be found as an example on the BUGS Website.

```
model bloodpressure

{for (i in 1:N)
   {for (t in 1:5)
      {y[i,t]~dnorm(ymean[i,t],tau.eps)
       ymean[i,t]<-p[i]+eta[i,t]+eta[i,t+1]
       }
    for (t in 1:6)
      {eta[i,t]~dnorm(0,tau.eta)
       }
    p[i]~dnorm(mu[group[i]],tau.p)
    }

 for (g in 1:2)
   {mu[g]~dnorm(mu0,0.001)
    }

 mu0~dnorm(150,0.0005)
 tau.eps~dgamma(1,30)
 tau.eta~dgamma(1,10)
 tau.p~dgamma(2,200)
 }
```

Figure 5.2: BUGS code for blood pressure example with moving average errors.

The weights of thirty young rats are measured at weekly intervals for five weeks. A straight-line model is used to relate weight to time. (We might well want to consider a more complicated form of curve and possibly allow autocorrelation of deviations from the curve but, for this example, we will stick to a straight line with independent "errors"). However the intercept and gradient of the line are allowed to vary as random effects between rats.

The five times, in days, at which the weights are measured are $t_1 = 8,\ t_2 = 15,\ t_3 = 22,\ t_4 = 29,\ t_5 = 36$. The weight of rat i on day t_j is

$$Y_{i,j} \mid \alpha_i, \beta_i, \tau_Y \sim N(\alpha_i + \beta_i[t_j - 22], \tau_Y^{-1}).$$

Now we need a model for how α_i, β_i vary between rats. We could simply write

$$\begin{aligned} \alpha_i \mid \mu_\alpha, \tau_\alpha &\sim N(\mu_\alpha,\ \tau_\alpha^{-1}) \\ \beta_i \mid \mu_\beta, \tau_\beta &\sim N(\mu_\beta,\ \tau_\beta^{-1}) \end{aligned} \tag{5.1}$$

with α_i, β_i independent given the parameters. However it might be more realistic to allow them to have a nonzero correlation. One way to do this (though not the way that it is done on the BUGS Website) is to specify the conditional distribution of β_i given α_i. So, instead of (5.1) we write

$$\beta_i \mid \mu_\beta, \tau_\beta, \alpha_i, \gamma \sim N(\mu_\beta + \gamma[\alpha_i - \mu_\alpha],\ \tau_\beta^{-1}).$$

Finally we give prior distributions to the model parameters. The priors given here are based (loosely) on those used in the example on the BUGS Website. They are meant to be "noninformative."

$$\begin{aligned} \mu_\alpha &\sim N(0,\ 10000) \\ \mu_\beta &\sim N(0,\ 10000) \\ \gamma &\sim N(0,\ 4000) \\ \tau_Y &\sim \text{Ga}(0.001,\ 0.001) \\ \tau_\alpha &\sim \text{Ga}(0.001,\ 0.001) \\ \tau_\beta &\sim \text{Ga}(0.001,\ 0.001) \end{aligned}$$

```
model rats

{for (i in 1:N)
   {for j in 1:5)
      {mean[i,j]<-alpha[i]+beta[i]*(t[j]-22)
       y[i,j] ~ dnorm(mean[i,j],tau.y)
       }
    alpha[i] ~ dnorm(mu.alpha,tau.alpha)
    betamean[i]<-mu.beta+gamma*(alpha[i]-mu.alpha)
    beta[i] ~ dnorm(betamean[i],tau.beta)
    }

 mu.alpha ~ dnorm(0,0.0001)
 mu.beta ~ dnorm(0,0.0001)
 gamma ~ dnorm(0,0.00025)
 tau.y ~ dgamma(0.001,0.001)
 tau.alpha ~ dgamma(0.001,0.001)
 tau.beta ~ dgamma(0.001,0.001)
 }
```

Figure 5.3: BUGS model specification for rats growth curves example.

Figure 5.3 shows suitable BUGS code.

Hospital	n_i	r_i	Hospital	n_i	r_i	Hospital	n_i	r_i
1	47	0	5	211	8	9	207	14
2	148	18	6	196	13	10	97	8
3	119	8	7	148	9	11	256	29
4	810	46	8	215	31	12	360	25

Table 5.2: Mortality in twelve hospitals performing cardiac surgery on babies. n_i : number of operations at hospital i. r_i : number of deaths at hospital i.

5.4 Practical 5

5.4.1 Hospital ranking

This example is taken from the BUGS Website. It concerns the mortality rates in twelve hospitals performing cardiac surgery in babies. The data are shown in table 5.2.

Crude methods of comparing hospitals might be misleading. For example, the variance of the observed proportions of deaths is large if the number of operations is smaller. Therefore a small hospital could appear to have a very bad rate simply because of a small number of cases. Using a random-effects model helps to smooth out such effects.

We suppose that, associated with hospital i there is a rate p_i which, if it were known, would be the probability of death at that hospital. We suppose that the number of deaths r_i out of n_i operations at hospital i has a binomial distribution

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

Then we write

$$b_i = \log\left(\frac{p_i}{1-p_i}\right)$$

and

$$b_i \mid \mu, \tau \sim N(\mu,\ \tau).$$

We then gives priors to the parameters. These are the priors used on the BUGS Website. They are so-called "noninformative" priors.

$$\begin{aligned} \mu &\sim N(0,\ 10^6) \\ \tau &\sim \text{Ga}(0.001,\ 0.001) \end{aligned}$$

1. Type the following model specification into a file called `hospitalbug.txt` .

```
model hospital

{for (i in 1:N)
   {r[i]~dbin(p[i],n[i])
    logit(p[i])<-b[i]
    b[i]~dnorm(mu,tau)
    }

 mu~dnorm(0.0,1.0E-6)
 tau~dgamma(0.001,0.001)
 }
```

   Note that `1.0E-6` means 1.0×10^{-6}.

2. Type the data into a file called `hospitaldata.txt` as follows.

```
list(N=12, n=c(47,148,119,810,211,196,148,215,207,97,256,360),
           r=c(0,18,8,46,8,13,9,31,14,8,29,24))
```

3. Use BRugs to evaluate the posterior distribution. Monitor $b_1, \ldots, b_{12}$ and compare the posterior 95% intervals for these. Does any hospital stand out from the rest?

   You will need to set some initial values. For example, to run two chains, create one file called

   `hospitalinits1.txt`

   containing the following

   ```
   list(mu=-2.0, tau=2.0)
   ```

   and another file called

   `hospitalinits2.txt`

   containing the following.

   ```
   list(mu=-2.0, tau=20.0)
   ```

   You would then need to issue commands as follows.

   ```
   modelCheck("hospitalbug.txt")
   modelData("hospitaldata.txt")
   modelCompile(2)
   modelInits("hospitalinits1.txt")
   modelInits("hospitalinits2.txt")
   modelGenInits()
   ```

   You would then be ready to start updating (with or without setting a monitor). The final `modelGenInits()` is necessary because our initial value files do not specify initial values for *all* of the unknowns.

5.4.2 Rat growth

This is the "rats" example of section 5.3.4. The BUGS code is available on the Web page as `ratsbug.txt` and there are two data files, also on the Web page, called `ratsxdata.txt` and `ratsydata.txt` . Because there are two data files, you will need to start like this.

```
modelCheck("ratsbug.txt")
modelData("ratsxdata.txt")
modelData("ratsydata.txt")
```

Use BRugs to evaluate the posterior distribution. Monitor μ_α, μ_β, γ, τ_y, τ_α and τ_β. You could also monitor the regression coefficients of individual rats, α_i, β_i, if you wish.

You will need to supply initial values for some of the unknowns. I suggest that you use two chains and initialise them as follows.

```
modelCheck("ratsbug.txt")
modelData("ratsxdata.txt")
modelData("ratsydata.txt")
modelCompile(2)
modelInits("ratsinits1.txt")
modelInits("ratsinits2.txt")
modelGenInits()
```

Here is a suggestion for the contents of `ratsinits1.txt`

```
list(tau.y=1.0E-5,tau.alpha=1.0,tau.beta=10.0)
```

and here is a suggestion for the contents of `ratsinits2.txt`

```
list(tau.y=5.0E-6,tau.alpha=10.0,tau.beta=100.0)
```

Set the sample monitors and do, say, 5000 updates. Then look at the results using `samplesHistory` . You might be surprised at how poor the convergence is in this example. This is probably because the parameters are poorly identified.

In fact things behave much better if we assume that α_i and β_i are conditionally independent given $\mu_\alpha, \mu_\beta, \tau_\alpha, \tau_\beta$. Try this. Replace the lines

```
betamean[i]<-mu.beta+gamma*(alpha[i]-mu.alpha)
beta[i] ~ dnorm(betamean[i],tau.beta)
```

with the single line

```
beta[i] ~ dnorm(mu.beta,tau.beta)
```

in the model file and try fitting the model again.

Further investigation shows that the posteriors for τ_Y, τ_α and τ_β are sensitive to the choice of priors for these parameters suggesting that the parameters are not well identified. Nevertheless, in this case, the "noninformative" priors seem to give sensible results.

5.4.3 Vertebral fractures in older women.

Here is one for you to do yourselves.

The data come from Cooper *et al.* (1991). Osteoporosis is a problem for many post-menopausal women. It can lead to bone fractures. Women were screened for evidence of vertebral fractures according to a certain criterion. A subset of the data were as follows.

i	Age group	Total number	Number with fracture
1	50-54	17	1
2	55-59	282	12
3	60-64	244	17
4	65-69	218	23
5	70-74	120	9
6	75-79	105	11
7	80-	18	5

Let the lower age limit of age-group i be x_i. Let the number of women screened in this group be n_i and let the number classified as having vertebral fractures be y_i. Then we assume

$$y_i \mid n_i, p_i \sim \text{Bin}(n_i,\ p_i)$$

with

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_1 x_i + \delta_i.$$

Here β_0 and β_1 are parameters about which we wish to learn. We adopt the following independent prior distributions.

$$\begin{aligned} \beta_0 &\sim N(-3,\ 5), \\ \beta_1 &\sim N(0,\ 1). \end{aligned}$$

Because the relationship between age and logit of fracture rate might not really be a straight line we allow some deviation by adding a random variable δ_i with

$$\delta_i \sim N(0, 0.001).$$

Use BRugs to evaluate the posterior distribution of $\beta_0,\ \beta_1,\ \eta_1, \ldots, \eta_7$.

Convergence and mixing are poor. You will need a long burn-in. It might help if you use two initial value files such as the following.

```
list(beta0=-4.5, beta1=0.01)

list(beta0=-5.5, beta1=0.06)
```

Try also monitoring $p_1, \ldots, p_7$. You can do this using

```
samplesSet("p")
```

The results are quite interesting.

References

Azzalini, A. and Bowman, A.W., 1990. A look at some data on the Old Faithful geyser. *Journal of the Royal Statistical Society, Series C*, **39**, 357-366.

Breslow, N.E. and Clayton, D.G., 1993. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Cooper, C., Shah, S., Hand, D.J., Adams, J., Compston, J., Davie, M. and Woolf,A., 1991. Screening for vertebral osteoporosis using individual risk factors. *Osteoporosis International*, **2**, 48-53.

Cowburn, G.J., 2003. *Bayesian mixture modelling with application to road traffic flow.* Ph.D. thesis, University of Sunderland.

Cowburn, G.J. and Farrow, M., 2007. Mixtures and diagnostic plots in modelling road traffic vehicle headways. Statistical Modelling, 7, 73-89.

Davies, O.L. and Goldsmith, P.L. (eds), 1972. *Statistical Methods in Research and Production*, 4th edition. Edinburgh: Oliver and Boyd.

Freeman, D.H., 1987. *Applied Categorical Data Analysis.* New York: Marcel Dekker.

Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M., 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972-85.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E., 1994. *A Handbook of Small Data Sets.* Chapman and Hall, London.

Krzanowski, W.J., 1988. *Principles of Multivariate Analysis.* Oxford: Oxford University Press.

Lunn, D.J., Thomas, A., Best, N.G., and Spiegelhalter, D.,J., 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325-337.

Open University, 1983. *MDST242 Statistics in Society, Unit C3: Is my child normal?.* Milton Keynes: The Open University.

Plummer, M., 2012. *JAGS Version 3.3.0 User Manual.*

Mosteller, F. and Tukey, J.W., 1977. *Data analysis and regression.* Reading, Massachusetts: Addison-Wesley.

Pearl,J., Geiger,D. and Verma,T., 1990, The logic of influence diagrams, In *Influence Diagrams, Belief Nets and Decision Analysis* (R.M.Oliver, J.Q.Smith eds.) New York: Wiley.

Phillips, D.P., 1978. Airplane accident fatalities increase just after newspaper stories about murder and suicide. *Science*, **201**, 748-750.

Scott, S.L., 2002. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337-351.

Snedecor, G.W. and Cochran, G.C., 1967. *Statistical Methods*, 6th edition. Iowa State University Press, Ames, Iowa.

Sokal, R.R. and Rohlf, F.J., 1981. *Biometry*, 2nd edition. San Francisco: W.H. Freeman.

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R., 1995. *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.50.* MRC Biostatistics Unit, Cambridge.

Thall, P.F. and Vail, S.C., 1990. Some covariance models for longitudinal count data with overdispersion. Biometrics, **46**, 657-71.

Till, R., 1974. *Statistical Methods for the Earth Scientist.* London: MacMillan.