6

Regression models

As in classical regression, Bayesian regression models are formulated by specifying a sampling distribution for the data (which we also loosely term the likelihood) and then a form of relationship between the assumed distribution of the response variable and any explanatory variables. The only difference is that we also specify prior distributions for the regression coefficients and any other unknown (nuisance) parameters. As we will see in this chapter, there are several advantages to a Bayesian approach, however, such as it being relatively straightforward to include parameter restrictions, use non-linear models, "robustify" against outliers, make predictions and inferences about functions of regression parameters, and handle missing data.

6.1 Linear regression with normal errors

Suppose our response variable is denoted y_i, $i = 1, ..., n$, and we have p covariates $x_{1i}, ..., x_{pi}$. We specify

$$y_i \sim \text{Normal}(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki},$$

along with prior distributions for $\beta_0, \beta_1, ..., \beta_p$ and σ. For example,

$$\beta_k \sim \text{Normal}(0, 100^2), \quad \log \sigma \sim \text{Uniform}(-100, 100)$$

or the alternative priors discussed in §5.2.4 and §5.2.7. Again we emphasise that if the specific choice of vague prior is influential, this suggests that a robust conclusion cannot be drawn from the data alone and more informative priors based on background information should be considered.

In Bayesian regression analysis it is generally advisable to consider "centering" any covariates, that is, subtracting the empirical mean from each value, as illustrated in the following example. This has the effect of reducing the posterior correlation between each coefficient $(\beta_1, ..., \beta_p)$ and the intercept term β_0, because the intercept is essentially relocated to the "centre" of the data. As discussed in §4.4, high levels of posterior correlation are problematic for Gibbs sampling.

Example 6.1.1. *Growth curve*
Gelfand et al. (1990) examine growth data from 30 young rats whose weights were measured weekly for five weeks. In this example we fit a linear regression to the 9th rat's data. The response variable y_i, $i = 1, ..., 5$, is the weight, in grams, on day x_i.

```
model {
  for (i in 1:5) {
    y[i]        ~ dnorm(mu[i], tau)
    mu[i]      <- alpha + beta*(x[i] - mean(x[]))
  }
  # Jeffreys priors
  alpha         ~ dflat()
  beta          ~ dflat()
  tau           <- 1/sigma2
  log(sigma2)   <- 2*log.sigma
  log.sigma     ~ dflat()
}

list(y = c(177,236,285,350,376), x = c(8,15,22,29,36))
```

We specify improper uniform priors for all parameters, and so the posterior mode will be equal to the maximum likelihood estimates: $\hat{\alpha} = 284.8$, $\hat{\beta} = 7.31$, $\hat{\sigma^2} = 71.3$ — note the posterior of σ^2 is extremely skewed. Figure 6.1 shows the posterior distribution of the model fit, produced through the `Inference->Compare` dialog box in WinBUGS.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	284.8	7.89	0.078	269.9	284.8	300.1	4001	10000
beta	7.316	0.7814	0.008582	5.82	7.316	8.819	4001	10000
sigma2	316.3	743.6	26.14	37.24	145.6	1586.0	4001	10000

Linear models where all or some of the covariates are categorical are sometimes called *analysis of variance* or *analysis of covariance* models, respectively, since the interest is often in comparing the variation of the outcome within and between categories. In BUGS these are treated just like any other linear regression — as a linear model with coefficients for each explanatory variable.

Example 6.1.2. *New York crime*
Press (1971) presents data on the effects of increasing police manpower in New York City. The response variable is the (seasonally adjusted) change in the number

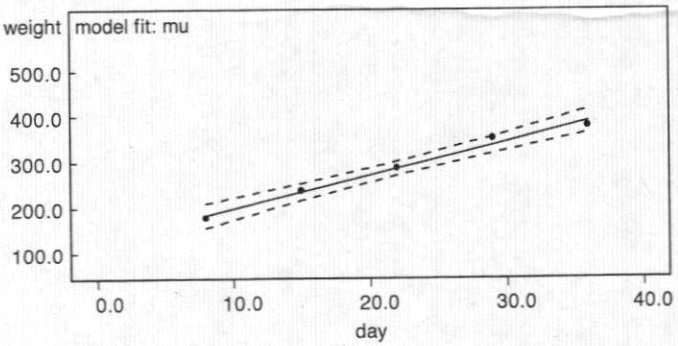

FIGURE 6.1
Model fit from Bayesian linear regression of rat 9's data in Example 6.1.1. The posterior median, 2.5% and 97.5% percentiles for each `mu[i]` are joined together by straight lines: the solid line joins the medians, whereas the 95% credible intervals are joined by dashed lines. The observed weights are shown by dots.

of thefts in 23 precincts of New York City from a 27-week base period in 1966 to a 58-week experimental period in 1966–1967. The percentage increase in manpower in each precinct (`MAN[]`) is also recorded, as is the district (`DIST[]`) to which each precinct belongs (1 = Downtown, 2 = Midtown, 3 = Uptown). The `DIST` covariate is a categorical variable and requires a slightly different approach to covariates that represent quantities. It doesn't make sense to include a term like `beta*DIST[i]` in the model because we can't realistically assume that the effect of going from downtown to midtown is the same as going from midtown to uptown, or even that they have the same sign. Instead, we can create and incorporate into the model two new covariates, one equal to one for midtown precincts, and zero otherwise (`D2[]`), and another equal to one for uptown precincts, and zero otherwise (`D3[]`). The model code for this *multiple regression* (where more than one covariate is included) is then

```
for (i in 1:23) {
  y[i]              ~ dnorm(mu[i], tau)
  D2[i]            <- equals(DIST[i], 2)
  D3[i]            <- equals(DIST[i], 3)
  mu[i]            <- beta0 + beta[1]*MAN[i]
                       + beta[2]*D2[i] + beta[3]*D3[i]
}
beta0               ~ dnorm(0, 0.0001)
for (j in 1:3) {
  beta[j]           ~ dnorm(0, 0.0001)
}
tau                <- 1/pow(sigma, 2)
# uniform prior on an interpretable scale
sigma               ~ dunif(0, 100)
```

Posterior summaries for the model parameters are shown below.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta[1]	-0.2378	0.1188	0.001181	-0.4759	-0.2374	-0.006211	5001	10000
beta[2]	0.558	3.027	0.03123	-5.493	0.5386	6.595	5001	10000
beta[3]	-4.03	3.13	0.03011	-10.14	-4.014	2.239	5001	10000
beta0	2.573	1.992	0.01811	-1.291	2.581	6.533	5001	10000
sigma	5.837	1.037	0.01537	4.243	5.686	8.233	5001	10000

Another way to implement the same model is to make use of BUGS' nested indexing feature. In this case we can make use of the DIST covariate directly via the following modification (note that there is no need to calculate D2 and D3 in this case):

```
mu[i] <- beta0 + beta[1]*MAN[i] + gamma[DIST[i]]
```

with

```
gamma[1] <- 0
gamma[2]  ~ dnorm(0, 0.0001)
gamma[3]  ~ dnorm(0, 0.0001)
```

where gamma[1] is fixed because only two district contrasts are identifiable — we could instead remove the intercept term, beta0, and estimate gamma[1] in its place. To provide initial values for a vector such as gamma, which contains both unknown parameters and constants (or logical nodes), we simply specify NA for any elements that are constant/logical, e.g., gamma = c(NA,0,0).

Typically in multiple regression problems, such as in Example 6.1.2 above, we are aiming for a parsimonious model. With this in mind we might wonder whether including a particular covariate in the model is worthwhile. Intuitively it may seem reasonable to require covariates appearing in the final model to have coefficients with high posterior probabilities of being non-zero. Informally we could say that a covariate effect is "significant" (at the 95% level) if the 95% posterior credible interval for the associated coefficient does not include zero. Credible intervals will vary as we include/remove different covariates in/from the model, and so, adopting this strategy, we are faced with the usual problems of forwards and backwards selection.

We will look at model criticism and comparison in more detail in Chapter 8. The Bayesian framework can actually accommodate situations in which the choice of covariates to be included in a given model is a model parameter itself — these methods are reviewed briefly in §8.8.2.

Example 6.1.3. *New York crime (continued)*

The credible intervals obtained in Example 6.1.2 above suggest that the effect of police manpower is "significant," whereas the district effects are not. Hence we might consider removing the D2 and D3 variables from the regression equation:

```
mu[i] <- beta0 + beta[1]*MAN[i]
```

Posterior summaries for `beta0`, `beta[1]`, and `sigma` are given below, and the resulting model fit is shown on the left-hand side of Figure 6.2. Note how the coefficients have changed values considerably, and that the effect of police manpower is no longer conventionally "significant," although it is close.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta[1]	-0.1761	0.1096	0.001085	-0.3921	-0.1765	0.04073	5001	10000
beta0	1.97	1.362	0.01394	-0.7489	1.963	4.632	5001	10000
sigma	5.873	0.9813	0.01054	4.297	5.74	8.136	5001	10000

6.2 Linear regression with non-normal errors

In classical linear modelling, the errors are usually assumed to be normally distributed, for example, the "least squares" estimators for linear regression are equivalent to maximum likelihood estimators under this assumption. However, we are not restricted to normality, and BUGS makes it easy to use any appropriate distribution. If we suspect outlying observations, for example, we can provide some robustness against their effects by assuming the data arise from a heavy-tailed t-distribution. Thus the outliers can be accommodated within the tails without necessarily forcing the location of the posterior to be moved significantly. The following example illustrates.

Example 6.2.1. *New York crime (continued): robust regression*

Note from the model fit shown on the left-hand-side of Figure 6.2 that the rightmost point is rather influential — without this point in place, positive and negative regression lines might seem equally plausible. The point corresponds to the 20th precinct (between the Hudson River and Central Park on the southwest side of Central Park). During the study, police manpower in the 20th precinct was *experimentally* increased by 40%, but no experimental changes were made in other precincts. Hence we might have cause to suspect that the corresponding observation could be an outlier. We robustify our analysis against the potential effects of such outliers with the following simple modification:

```
y[i] ~ dt(mu[i], tau, dof)
```

where `dt(x,y,z)` denotes a Student-t distribution with mean `x`, precision parameter `y` and degrees of freedom `z` (see Appendix C.1 — note the variance is $\frac{z}{y(z-2)}$ for $z > 2$). In principle, we could estimate the degrees of freedom, as in Example 4.1.2, by assigning an appropriate prior, but this can be problematic unless there are many observations. Instead, here, we set `dof <- 4` to give a very heavy

tailed distribution for the residuals. The model fit is shown on the right-hand side of Figure 6.2 and posterior summaries are given in the table below:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	1.699	1.632	0.02879	-1.191	1.59	5.222	5001	10000
beta[1]	-0.1244	0.1449	0.002557	-0.358	-0.1455	0.2097	5001	10000
sigma	4.883	1.035	0.01803	3.266	4.756	7.214	5001	10000

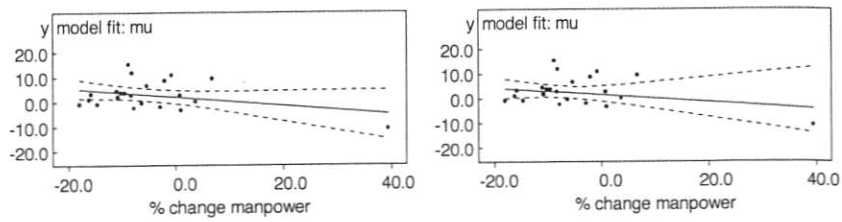

FIGURE 6.2
Model fits for New York crime data in Examples 6.1.3 and 6.2.1 with manpower effect alone in the regression equation: left-hand side, normal residuals; right-hand side, t_4 distributed residuals.

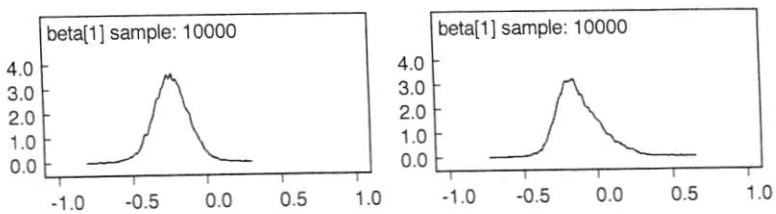

FIGURE 6.3
Posterior density estimates for manpower effect in regression analysis of New York crime data in Examples 6.1.3 and 6.2.1: left-hand side, normal residuals; right-hand side, t_4 distributed residuals.

Note that `sigma` is no longer the residual standard deviation; this is given by `sd <- sigma*sqrt(dof/(dof-2))`, which is `sigma*1.414` in this case. Hence the posterior median residual standard deviation is 6.73, a little higher than before. Just looking at the posterior median model fit we might think that assuming a t distribution for the residuals has had a negligible effect. However, note that the posterior summaries show the effect of manpower has been attenuated, with the posterior median for `beta[1]` reduced in size from -0.177 (with normal errors) to

−0.146. This is consistent with less of the overall variability apparent in the data being explained by the model, as indicated by the increased residual standard deviation. Also note that use of the t distribution may affect our substantive inferences, since we can no longer consider the effect of manpower "significant" — considerably more posterior probability now lies to the right of zero. This is reflected by the (now positive) slope of the upper end of the credible interval for the model fit, and by the posterior density estimates for `beta[1]` shown in Figure 6.3.

6.3 Non-linear regression with normal errors

MCMC methods can easily accommodate non-linear regressions. The only additional effort required in fitting such models might be ensuring that the parameters always have meaningful values by imposing appropriate constraints. This is illustrated in the following example. Note that we can easily extend non-linear models to non-normal errors, as in the previous subsection.

Example 6.3.1. *Dugongs*
Carlin and Gelfand (1991) consider data on length (y_i) and age (x_i) measurements for $i = 1, ..., n = 27$ dugongs (sea cows) captured off the coast of Queensland. The data are shown in Figure 6.5. A frequently used nonlinear growth curve with no inflection point and an asymptote as $x \to \infty$ is the Von Bertalanffy growth model, given by

$$y_i \sim \text{Normal}(\mu_i, \sigma^2), \quad \mu_i = L_\infty - (L_\infty - L_0)\, e^{-K x_i}, \quad i = 1, ..., n,$$

where $L_\infty > L_0 > 0$ and $K > 0$. L_∞ represents the maximum expected length achievable, and L_0 is the length at time 0. We can impose such constraints in various ways, e.g., $L_0, K \sim \text{Uniform}(0, 100)$, $L_\infty = L_0 + \beta, \beta \sim \text{Uniform}(0, 100)$. We illustrate the use of this particular prior in the code below (model 1). We also illustrate the use of truncated normal priors for L_∞ and L_0 (model 2). We can use the `I(,)` syntax to represent truncated distributions, as discussed in §9.6 and Appendix A.2.2, as there are no unknown parameters in the prior distribution. In addition, we present two further constrained priors based on the fact that the von Bertalanffy model can be rewritten as $\mu_i = \alpha - \beta\gamma^{x_i}$, with $\alpha = L_\infty > 0$, $\beta = L_\infty - L_0 > 0$, and $0 < \gamma = e^{-K} < 1$. In model 3 we assume $\alpha, \beta \sim \text{Uniform}(0, 100)$ and $\gamma \sim \text{Uniform}(0, 1)$. Model 4 is the same as model 3 except that we use approximately the same prior for γ as in model 1, by assuming $\gamma \sim \text{Gamma}(0.001, 0.001)I(0, 1)$ (since $p(K) \propto 1$ is equivalent to $e^{-K} \sim \text{Gamma}(0, 0)I(0, 1)$). Four copies of the data are supplied and we compare posterior distributions for α, β, γ, and σ^2 between the four priors.

```
model {
  for(j in 1:N) {
    for (i in 1:4) {
      y[i,j]                  ~ dnorm(mu[i,j], tau[i])
    }
    mu[1,j] <- Linf[1] - (Linf[1] - L0[1])*exp(-K[1]*x[1,j])
    mu[2,j] <- Linf[2] - (Linf[2] - L0[2])*exp(-K[2]*x[2,j])
    mu[3,j] <- alpha[3] - beta[3]*pow(gamma[3], x[3,j])
    mu[4,j] <- alpha[4] - beta[4]*pow(gamma[4], x[4,j])
  }
  L0[1]                        ~ dunif(0, 100)
  L0[2]                        ~ dnorm(0, 0.0001)I(0, Linf[2])
  Linf[1]                     <- L0[1] + beta[1]
  Linf[2]                      ~ dnorm(0, 0.0001)I(L0[2], )
  K[1]                         ~ dunif(0, 100)
  K[2]                         ~ dunif(0, 100)
  for (i in 1:2) {alpha[i] <- Linf[i]}
  for (i in 3:4) {alpha[i]  ~ dunif(0, 100)}
  beta[1]·                     ~ dunif(0, 100)
  beta[2]                     <- Linf[2] - L0[2]
  for (i in 3:4) {beta[i]   ~ dunif(0, 100)}
  for (i in 1:2) {gamma[i] <- exp(-K[i])}
  gamma[3]                     ~ dunif(0, 1)
  gamma[4]                     ~ dgamma(0.001, 0.001)I(0, 1)
  for (i in 1:4) {
    tau[i]                    <- 1/sigma2[i]
    log(sigma2[i])            <- 2*log.sigma[i]
    log.sigma[i]               ~ dunif(-10, 10)
  }
}
```

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha[1]	2.65	0.07281	0.001407	2.527	2.644	2.809	10001	50000
alpha[2]	2.651	0.07263	0.001245	2.529	2.644	2.814	10001	50000
alpha[3]	2.656	0.07748	0.001929	2.532	2.647	2.829	10001	50000
alpha[4]	2.654	0.07424	0.001737	2.528	2.647	2.819	10001	50000
beta[1]	0.9751	0.07746	0.001512	0.8275	0.9736	1.129	10001	50000
beta[2]	0.9747	0.07807	6.402E-4	0.8263	0.9733	1.129	10001	50000
beta[3]	0.9759	0.07796	0.001022	0.828	0.9744	1.135	10001	50000
beta[4]	0.9759	0.07727	9.129E-4	0.8288	0.9742	1.132	10001	50000
gamma[1]	0.8607	0.03351	7.849E-4	0.7833	0.8646	0.9146	10001	50000
gamma[2]	0.8613	0.03373	5.879E-4	0.7845	0.8651	0.9161	10001	50000
gamma[3]	0.8632	0.03293	7.05E-4	0.7892	0.8665	0.9189	10001	50000
gamma[4]	0.8623	0.03386	6.953E-4	0.7839	0.8662	0.917	10001	50000
sigma2[1]	0.009987	0.003213	2.702E-5	0.005568	0.009403	0.01791	10001	50000
sigma2[2]	0.009961	0.003191	2.136E-5	0.005532	0.009387	0.01774	10001	50000
sigma2[3]	0.009973	0.003169	2.552E-5	0.005552	0.009384	0.01777	10001	50000
sigma2[4]	0.009975	0.003194	2.505E-5	0.005582	0.009389	0.01786	10001	50000

The results are virtually identical for all four prior distributions, even though the priors themselves differ considerably, as illustrated in Figure 6.4. The model fit is shown in Figure 6.5.

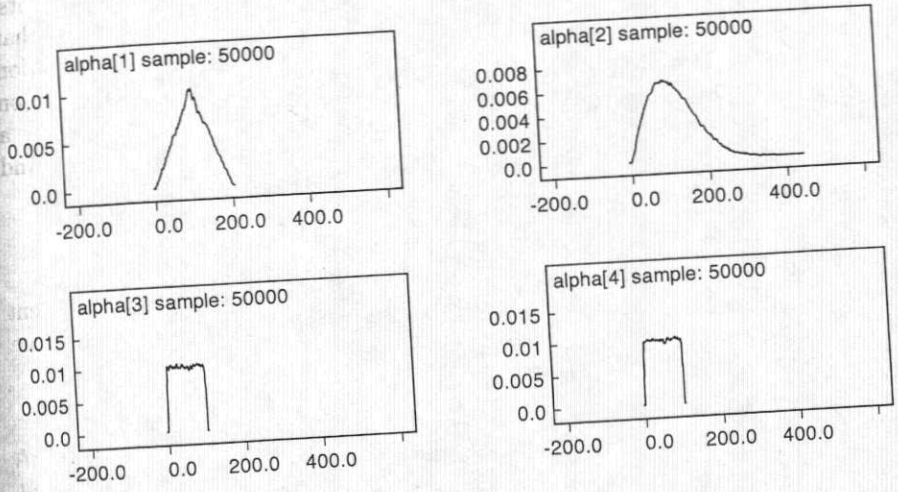

FIGURE 6.4
Prior distributions for α used in dugongs analyses — Example 6.3.1.

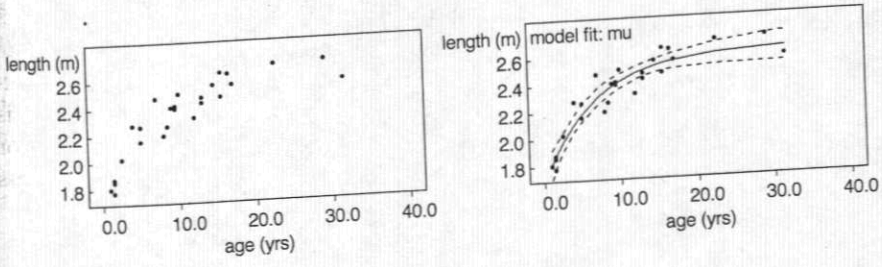

FIGURE 6.5
Left-hand side: dugong lengths (m) plotted against age in years. Right-hand side: same data with accompanying model fit.

6.4 Multivariate responses

Suppose we have observed n measurements on each of a number of individuals. Let i $(= 1, ..., N)$ index individuals and j $(= 1, ..., n)$ index measurements, and let Y_{ij} denote the jth observation made on individual i. Suppose also that measurements are made on the whole real line, so that a normality assumption might be appropriate. To account for the fact that observations made on the same individual may be correlated, we can assume that they follow a multivariate normal (MVN) distribution with unknown mean vector μ and variance-covariance matrix Σ. That is,

$$Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})' \sim \text{MVN}_n(\mu, \Sigma), \quad i = 1, ..., N.$$

If we have also observed covariates, such as the age at which each measurement was taken, specification of appropriate forms for the elements of μ leads to a multivariate regression model:

$$\mu_j = \beta_0 + \sum_{k=1}^{p} \beta_k x_{kj}, \quad j = 1, ..., n,$$

where x_{kj} denotes the jth value of covariate k. Typically we would specify vague normal priors for the coefficients, i.e. $\beta_. \sim \text{Normal}(0, 100^2)$, and an inverse-Wishart prior (see Appendix C.4) for the covariance Σ, via $\Sigma^{-1} \sim \text{W}(R, \rho)$. Here the right-hand side denotes a Wishart distribution with "scale matrix" R and degrees of freedom ρ. The Wishart distribution is the multivariate analogue of the gamma distribution and arises in classical statistics as the distribution of the sum-of-squares-and-products matrix in multivariate normal sampling. It is the conjugate prior for the precision matrix of a multivariate normal distribution. The least informative, proper Wishart prior is given by setting $\rho = p$, where p is the dimension of the distribution. The prior mean is ρR^{-1} and so a good choice for R is $\rho\Sigma_0$, where Σ_0 is some prior guess for the covariance.

Example 6.4.1. *Jaws*

Elston and Grizzle (1962) present repeated measurements of jawbone height on 20 boys. Each boy's jawbone was measured at ages 8, 8.5, 9, and 9.5 years, and interest focuses on describing the average growth curve of the jawbone. BUGS code for a multivariate regression model is given below.

```
model {
  for (i in 1:20) {Y[i, 1:4] ~ dmnorm(mu[], Sigma.inv[,])}
  for (j in 1:4)  {mu[j]     <- alpha + beta*x[j]}
  alpha                      ~ dnorm(0, 0.0001)
  beta                       ~ dnorm(0, 0.0001)
```

```
  Sigma.inv[1:4, 1:4]          ~ dwish(R[,], 4)
  Sigma[1:4, 1:4]             <- inverse(Sigma.inv[,])
}

list(Y = structure(
        .Data = c(47.8, 48.8, 49.0, 49.7,
                  46.4, 47.3, 47.7, 48.4,
                     ..............
                  46.3, 47.6, 51.3, 51.8),
        .Dim =  c(20, 4)),
     x = c(8.0, 8.5, 9.0, 9.5),
     R = structure(
        .Data = c(4, 0, 0, 0,
                  0, 4, 0, 0,
                  0, 0, 4, 0,
                  0, 0, 0, 4),
        .Dim =  c(4, 4)))
```

Array quantities in BUGS, such as `Sigma.inv`, must have their dimensions (`1:4`, `1:4` in this case) specified when they are defined, but not when they are used in the definitions of other nodes. See Appendix A.5. Note the use of the `structure()` syntax to specify data in matrix format — the data for `R` is supplied as a vector formed by concatenating successive rows of the matrix — see §12.4.2. The value of `R` is set equal to $\rho\Sigma_0$ where $\Sigma_0 = I$ is chosen by guessing the order of magnitude of variation between responses. Peeking at the data in order to set the prior is generally inappropriate as it is, strictly speaking, using the data twice in the analysis. However, assessing the order of magnitude of the variability is reasonable.

Also note that the multivariate normal distribution (`dmnorm`) in BUGS follows its univariate counterpart in being parameterised in terms of precision (Σ^{-1}). The matrix-valued `inverse()` function then allows inference on Σ. The model fit is shown in Figure 6.6 below. As an alternative for these data, we could have used a hierarchical "random coefficients" model — see Chapter 10.

Multivariate linear regressions are easily extended to nonlinear regressions, as in the univariate case. In addition, we can also specify a multivariate t-distribution (`mvt`) for the errors, to accommodate any outlying individuals. One area in which we do not have much freedom, however, is with the Wishart prior. Covariance matrices must always be positive-definite in order for them to make sense. The Wishart distribution is the only standard distribution that imposes this constraint naturally. If we wish to use an alternative form of prior then we must take responsibility for imposing the constraint ourselves, through appropriate parameterisation, say — the software will almost certainly crash if the constraint is not satisfied. Alternative priors for covariance matrices are further discussed in § 10.2.3 and in Gelman et al. (2004),

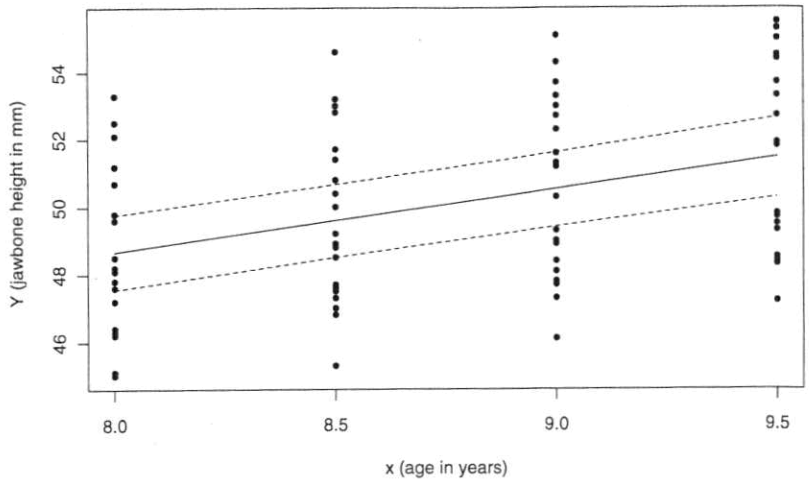

FIGURE 6.6
Model fit for jawbone data. The posterior median fit and 95% credible interval are indicated by the solid and dashed lines, respectively.

p. 483. Multivariate normal distributions with two specific structural forms for the covariance matrix are implemented as separate distributions in the BUGS language — see §11.3.6 for details.

6.5 Generalised linear regression models

Specification of Bayesian generalised linear models (GLMs) follows straightforwardly from the above discussion of linear models. No closed-form solution is available, but as we have seen for nonlinear models, it is still straightforward to obtain posterior samples using MCMC. The main differences with GLMs are that the sampling distribution of the data is typically non-normal and that we use a "link function" to transform parameters of that distribution onto a scale where a linear model can be used appropriately. More formally we assume that the data, y_i, $i = 1, ..., n$, arise from a specific distribution in the exponential family (McCullagh and Nelder, 1989), with

$$E[y_i] = \mu_i = g^{-1}(\eta_i), \quad \eta_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki},$$

for covariates x_{ki}, $k = 1, ..., p$. The exponential family of distributions includes distributions such as normal, Poisson, and binomial. Appropri-