Bayesian GLMs in R for Ecology

Mark Warren & Carl Smith

November 2021

Contents

7

0.1 Getting started with R and RStudio 7

1 Introduction to Bayesian inference 9

1.1 The differences between Bayesian and frequentist approaches 11

1.2 Fitting Bayesian GLMs 11

1.3 Priors 11

1.4 The posterior distribution 11

1.5 Bayesian computational methods 11

1.6 The advantages of Bayesian inference 11

1.7 Criticism of Bayesian inference 11

1.8 Conclusions 11

2 Data exploration 13

2.1 Six-step data exploration protocol 14

2.2 Results of data exploration 14

2.3 Conclusions 14

3 Bayesian GLM 15

3.1 European bitterling territoriality 16

3.2 Steps in fitting a Bayesian GLM 17

3.3 Conclusions 56

4 Bayesian Poisson GLM 57

4.1 Stickleback lateral plate number 59

4.2 Steps in fitting a Bayesian GLM 59

4.3 Conclusions 60

5 Bayesian negative binomial GLM 61

5.1 Coral abundance 63

5.2 Steps in fitting a Bayesian GLM 63

5.3 Conclusions 63

6 Bayesian Bernoulli GLM 65

6.1 Common cuckoo parasitism of great reed warbler nests . 67

6.2 Steps in fitting a Bayesian GLM 67

6.3 Conclusions 67

7 Bayesian gamma GLM 69

7.1 Common seal dive duration 71

7.2 Steps in fitting a Bayesian GLM 71

7.3 Conclusions 71

8 Implementing and assessing Bayesian GLMs 73

8.1 Prior information 74

8.2 Presenting the results of a Bayesian GLM 74

8.3 Reviewing Bayesian GLMs 74

8.4 Misuse of Bayesian inference 74

8.5 Conclusions 74

9 Coda 75

Placeholder

0.1 Getting started with R and RStudio

0.1.1 Basic points

0.1.2 Navigating RStudio

0.1.3 Basic settings in RStudio

0.1.4 Basic principles in R

0.1.5 Working with RStudio projects

0.1.5.1 Creating projects

0.1.5.2 Opening and closing projects

0.1.5.3 Working with multiple projects

0.1.6 Importing data

0.1.7 Functions and packages

Chapter 1

Introduction to Bayesian inference

Placeholder

1.1 The differences between Bayesian and frequentist approaches

1.1.1 Frequentist approach

1.1.2 Bayesian approach

1.1.3 Bayes' theorum

1.1.4 A frequentist or Bayesian framework?

1.2 Fitting Bayesian GLMs

1.2.1 Steps in fitting a Bayesian GLM

1.3 Priors

1.3.1 Non-informative priors

1.3.2 Weakly-informative priors

1.3.3 Informative priors

1.3.4 Conjugate priors

1.4 The posterior distribution

1.5 Bayesian computational methods

1.5.1 Markov chain Monte Carlo sampling (MCMC)

1.5.2 Numerical approximation

1.6 The advantages of Bayesian inference

1.7 Criticism of Bayesian inference

Chapter 2

Data exploration

Placeholder

2.1 Six-step data exploration protocol

2.1.1 Outliers

2.1.2 Normality and homogeneity of the dependent variable

2.1.3 Lots of zeros in the response variable

2.1.4 Multicollinearity among covariates

2.1.5 Relationships among dependent and independent variables

2.1.6 Independence of response variable

2.2 Results of data exploration

2.3 Conclusions

Chapter 3

Bayesian GLM

A General Linear Model predicts a dependent (or response) variable, which is continuous and approximately normally distributed, from one or more independent (or predictor) variables. A normal statistical distribution, also referred to as a Gaussian distribution (after the brilliant German mathematician Carl Friedrich Gauss), assumes the data are drawn from a distribution that is symmetric and can be summarised by the arithmetic mean and standard deviation. Independent variables may also be continuous, categorical, or a combination of continuous and categorical. Categorical variables are commonly referred to as *factors*, which have a series of *levels.* For example, a factor might be sex, which has just two levels (male and female).

A GLM comprises three components: 1. the linear predictor, which is a linear function of the predictor variable; 2. the conditional probability distribution of the response variable, which is the distribution of the response variable across the regression line for the given set of predictor variables; 3. the link function, which connects the linear predictor with the mean of the conditional probability distribution.

Choice of conditional probability distribution (such as Gaussian, binomial, Bernoulli, Poisson, gamma, beta, etc.) is not based on the distribution of the raw response variable, but rather on variable characteristics, such as whether the variable is continuous or discrete, bounded or unbounded. Choice of conditional distribution largely determines which link function is most appropriate (such as identity, log, logit, inverse, etc.), though choice of link function can be refined as part of the model fitting process.

3.1 European bitterling territoriality

In this Chapter we fit a Bayesian General Linear Model with a Gaussian conditional distribution and an identity link function to a set of data on male European bitterling (*Rhodeus amarus*) territorial behaviour. Bitterling are small freshwater fish with an unusual breeding system. During the breeding season, male bitterling are aggressively territorial and guard freshwater mussels. Female bitterling develop a long egg-laying tube ('ovipositor') that they use to place their eggs in the gills of the mussel, which the males then fertilise.

A study was conducted in Lake Dědová near Lanžhot in the Czech Republic to measure the response distance of male bitterling to a rival when they were guarding a mussel. Male response distance was measured by gradually moving a model of a male bitterling towards a territorial male that was guarding a mussel. The response distance was the horizontal distance that it was possible to move the model towards a guarded mussel before the territorial male attacked it. After obtaining an estimate of the response distance, territorial males were captured with a hand net and their length measured, after which they were immediately released.

In addition, males were randomly allocated to a food supplement treatment, with approximately half the males in the study receiving a 1 g cube of freeze-dried *Tubifex* worms daily for six days before the start of data recording. The remaining males received no food supplement, but did experience disturbance each day that was comparable to those receiving a food supplement.

A two-day pilot study with 8 individuals was also conducted. Data from the pilot study were used to assign prior distributions to fixed parameters in the model.

Import data

Data for European bitterling are saved in a comma-separated values (CSV) file `bitterling.csv` and are imported into a dataframe in R using:

```
bitt <- read_csv("bitterling.csv")
```

Start by inspecting dataframe bitt:

```
str(bitt)
```

```
spec_tbl_df [48 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ male     : num [1:48] 1 2 3 4 5 ...
 $ sl       : num [1:48] 37 58 50 52 42 ...
 $ supp_feed: num [1:48] 0 0 0 1 0 ...
 $ resp_dist: num [1:48] 114 172 168 185 135 ...
 - attr(*, "spec")=
  .. cols(
  ..   male = col_double(),
  ..   sl = col_double(),
  ..   supp_feed = col_double(),
  ..   resp_dist = col_double()
  .. )
```

The dataframe comprises 48 observations of 4 variables. Each row in the dataframe represents an observation for a different male bitterling (male). The variable sl is continuous and represents the standard length (in mm) of each male bitterling, while the variable supp_feed is categorical (though coded numerically) indicating those males that received no food supplement (0) and those that did (1). The variable resp_dist is the aggressive response distance (in cm) and is the response (dependent) variable of interest.

3.2 Steps in fitting a Bayesian GLM

We will follow the 9 steps to fitting a Bayesian GLM, detailed in Chapter 2:

1. *State the question*
2. *Perform data exploration*
3. *Select a statistical model*
4. *Specify and justify a prior distribution on parameters*
5. *Fit the model*

6. *Obtain the posterior distribution*
7. *Conduct model checks*
8. *Interpret and present model output*
9. *Visualise the results*

3.2.1 State the question

This study was conducted to understand the extent to which the territorial behaviour of male bitterling is a function of male size and body condition. Our predictions were that larger males would be more effective in responding to intruders to their territory than smaller males. A further prediction was that, given that territoriality is energetically expensive, and males are often constrained in their feeding while engaged in territory defence, supplementing the diets of males would also increase the aggressive response distance of males. A final prediction was that these two variables would interact; specifically that the effect of body size on response distance would be less pronounced in males that received a food supplement; i.e. energy depletion is more severe for larger males.

Consequently there are three specific predictions to test:

1. A positive association between male body size, measured as standard length (`sl`), and response distance (`resp_dist`).
2. A positive association between provision of supplementary food (`supp_feed`) and response distance.
3. An interaction between male body size and supplementary feeding and response distance, with a steeper slope between body size and response distance for males that did not receive supplementary food.

3.2.2 Data exploration

As with any analysis, whether Bayesian or frequentist, we start by conducting a data exploration to identify any potential problems with the data. First check for missing data.

```
colSums(is.na(bitt)
```

```
     male        sl supp_feed resp_dist 
        0         0         0         0 
```

No missing data.

3.2.2.1 Outliers

Outliers in the data can identified visually using multi-panel Cleveland dotplots (R code is available in the R script associated with this chapter):

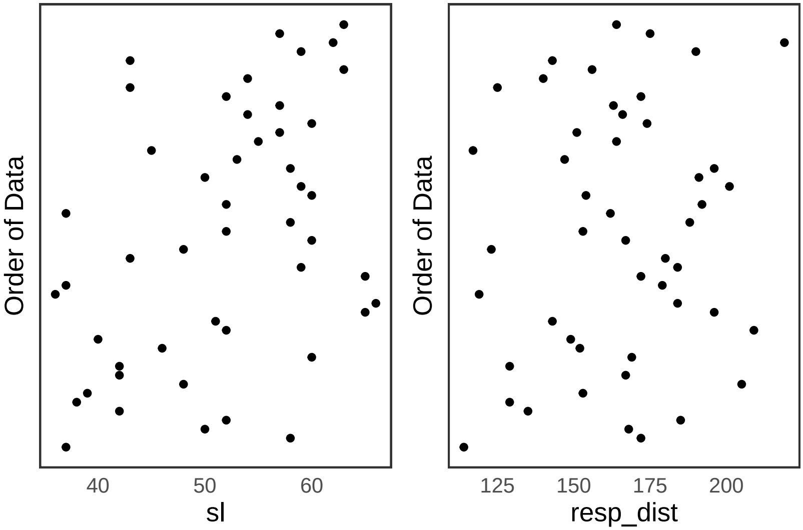

Figure 3.1: **Dotplots of male standard length (mm) and aggressive response distance (cm) of European bitterling. Data are arranged by the order they appear in the dataframe.**

There are no outliers in Fig. 3.1.

3.2.2.2 Normality and homogeneity of the dependent variable

An assumption of a Bayesian Gaussian GLM is that the response variable is normally distributed at each level of the covariate values. The distribution of a continuous variable can be visualized by dividing the x-axis into “bins” and counting the number of observations in each bin as a frequency polygon using the `geom_freqpoly()` function from the `ggplot2` package:

```
bitt %>%'  ggplot(aes(resp_dist)) +  geom_freqpoly( bins =
6) +  labs(x = "Response distance (cm)", y = "Frequency") +
My_theme
```

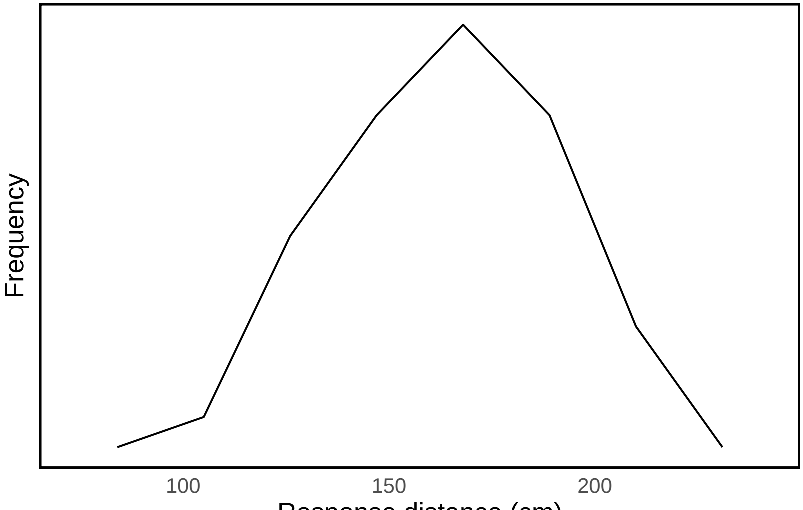

Figure 3.2: **Frequency polygon of response distance (cm) of male European bitterling to the model of a rival.**

The frequency polygon plot of the dependent variable (Fig. 3.2) shows a distribution that looks approximately normal.

3.2.2.3 Balance of categorical variables

The categorical variable for the supplementary feeding treatment (`supp_feed`) is coded numerically (0 = no supplementary feeding, 1 = supplementary feeding). This variable needs to be designated as a factor.

```
bitt$Supp <- factor(bitt$supp_feed)
```

We then examine the balance of this variable:

```
table(bitt$fSupp)
```

```

 0  1 
25 23 
```

Balance is not perfect, with 25 males in the no supplementary feeding treatment and 23 receiving supplementary feeding, but the balance is acceptable.

3.2.2.4 Multicollinearity among covariates

Along with normality of residuals and homogeneity of variance, an additional assumption of linear modelling is independence of the independent variables. In ecological studies it is not unusual to collect a large number of variables, which are often highly correlated. If covariates in a model are correlated, then the model may produce unstable parameter estimates with inflated standard errors.

Multicollinearity can be tested in several ways. We can obtain a comprehensive summary of the relationship between the two model covariates using the `ggpairs` command from the `GGally` package:

```
bitt %>% ggpairs(columns = c("sl", "fSupp"), aes(colour =
fSupp, alpha = 0.8), lower = list(continuous = "smooth_loess",
combo = wrap("facethist", binwidth = 5))) + My_theme
```

The plot matrix in Fig. 3.3 demonstrates no clear pattern of collinearity between the two covariates and illustrates good overlap in male standard length between levels of the (randomly assigned) feeding treatment.

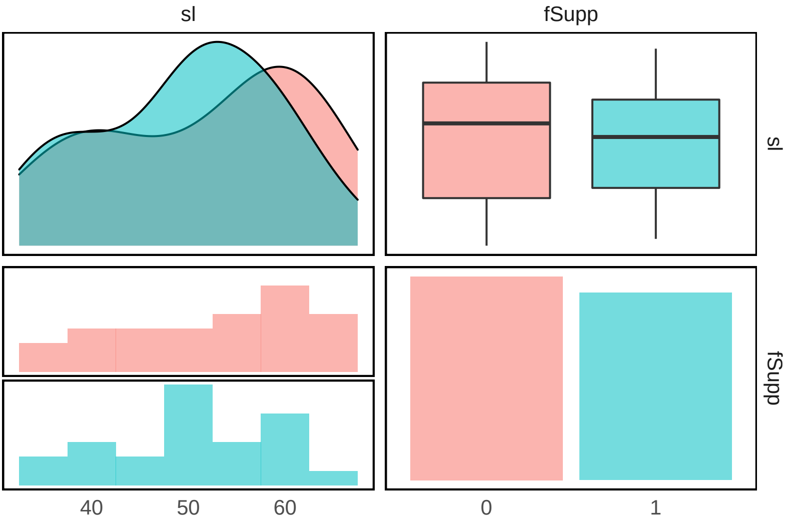

Figure 3.3: **Plot matrix of bitterling standard length (mm) and supplementary feeding treatment. The top left panel shows a frequency plot of standard length split by feeding treatment, while the top right shows the same data expressed as a box-plot. The lower left panel shows a length-frequency histogram of standard lengths, with those for males that did not receive supplementary feeding above and those that did, below. The lower right panel shows the total number of individual males in each supplementary feeding treatment.**

Another approach to identifying multicollinearity is by calculating a variance inflation factor (VIF) for each variable. The VIF is an estimate of the proportion of variance in one predictor explained by all the other predictors in the model. A VIF of 1 indicates no collinearity. VIF values above 1 indicate increasing degrees of collinearity. VIF values exceeding 3 are considered problematic (Zuur et al., 2009), in which case the variable with the highest VIF should be removed from the model and the VIFs for the model recalculated.

The VIF for a model can be estimated using the `vif` function from the `car` package:

```
round(vif(lm(resp_dist ~ sl + fSupp, data = bitt)),2)
```

1.01, 1.01

For the bitterling model the estimated VIFs are <3, so there is no problem with multicollinearity.

3.2.2.5 Zeros in the response variable

Zeros should not be omitted from a dataset. However, an excess of zeros in the response variable, termed ‘zero inflation,’ can cause problems with an analysis. The number of zeros in the response variable can be calculated with:

```
sum(bitt$ resp_dist == 0)
```

```
[1] 0
```

There are no zeros in the response variable, indicating that all territorial males responded aggressively to intruders.

3.2.2.6 Relationships among dependent and independent variables

Visual inspection of the data using plots is a critical step and will illustrate whether relationships are linear or non-linear and whether there are interactions between covariates. R code for this plot is available in the R script associated with this chapter.

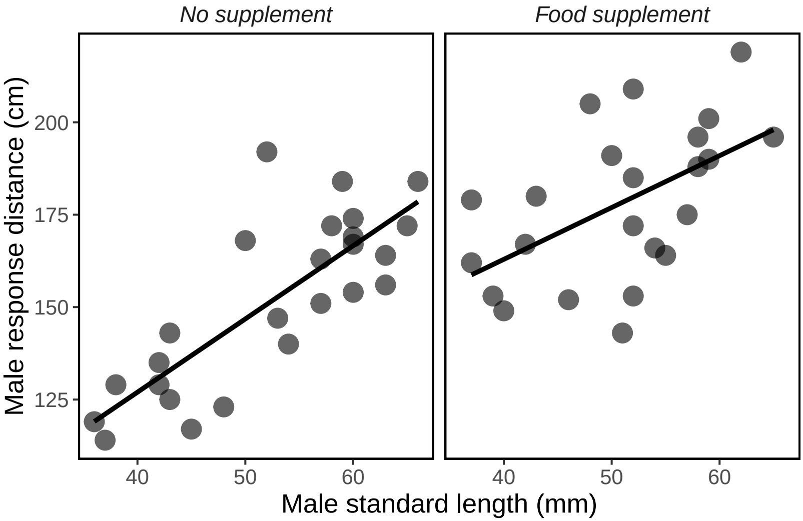

Figure 3.4: **Multipanel scatterplot of male response distance (cm) and standard length (mm) of European bitterling either without or receiving supplementary feeding with a line of best fit plotted.**

The plot of the data in Fig. 3.4 does not indicate a non-linear pattern in the data. However, fitted lines for the relationship between male response distance (cm) and standard length (mm) do suggest that the nature of this relationship may vary with feeding treatment, implying a potential interaction between fish size and feeding treatment; i.e. the slopes differ between treatments. An interaction would mean that the relationship between response distance and standard length depends on nutritional state. Interactions like this one are biologically interesting. Given the pattern in these data, inclusion of an interaction term in the model is justified.

3.2.2.7 Independence of response variable

An assumption for a GLM is that each observation in a dataset is independent of all others. In the case of the present study each row of data was a different male bitterling. The study was conducted over a short period (10 days) at the peak of the spawning season of the species in a single lake, which reduced the risk of any strong temporal and spatial effects. Observations were also made by a single experimenter, limiting the risk of dependency in the data due to variation in observer bias. On this basis, we will assume the data are independent.

3.2.3 Selection of a statistical model

The study was designed specifically to understand the extent to which the territorial behaviour of male European bitterling is a function of male size and nutritional state. The dependent variable is male response distance, which the data exploration showed to be continuous and approximately normally distributed (Fig. 3.2). There are no zeros in the response variable and there is good reason to believe data are independent. The relationship between male standard length and response distance is approximately linear, irrespective of food supplementation (Fig. 3.4).

Given these findings, a Gaussian is an appropriate distribution as a starting point, in combination with an *identity* link function (essentially no link function). Two covariates will be included in the model; male standard length (continuous) and food supplementation (categorical, with two levels) as well as their interaction, which means the model will have five parameters; an intercept (β_1), a slope for standard length (β_2), food supplementation (β_3), and the interaction between standard length and food supplementation (β_4), and the variance (σ^2) of the normal distribution for male response distance.

In the context of an INLA model, the variance parameter is termed a *hyperparameter*. In a simple linear model the hyperparameter just comprises the model residual variance. However, in more complex models the hyperparameter may also include other variance components, such as the random effects in a mixed model or the smoother in a Generalised Additive Model (GAM).

For computational efficiency, Bayesian analysis uses the precision (τ or tau) of parameters rather than variance. Precision is the reciprocal of the variance (σ^2), thus:

$\tau = \sigma^{-2}$

By default a diffuse gamma prior is assumed for the precision in INLA.

3.2.4 Specification of priors

A key aspect of any Bayesian model are the priors placed on model parameters. While there has been a tendency by ecologists to use non-informative or weakly informative priors, carefully formulated informative priors offer a powerful approach to modelling data, taking the modelling process beyond a description of the data and incorporating additional data or previous findings in a model (see Chapter 2).

3.2.4.1 Pilot study

In the study described here, a 2-day pilot experiment was conducted before the main study. This pilot study provided an opportunity for refining data collection methods and to obtain model priors. A total of 8 males were tested in the pilot experiment, with 4 receiving a food supplement and 4 with no supplement. In the pilot study several alternative food supplements were used, which meant the protocol followed was not identical to the main study, though the findings broadly matched the observations from the main study.

Import pilot data

Pilot data are saved in the tab-delimited file pilot.txt and are imported into a dataframe in R using the command:

```
pilot <- read_tsv("pilot.txt")
```

Note we use the `read_tsv()` function from the `readr` package which is part of the `tidyverse` set of packages.

Start by inspecting the dataframe:

```
str(pilot)
```

```
spec_tbl_df [8 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ order     : num [1:8] 1 2 3 4 5 6 7 8
 $ supplement: chr [1:8] "no" "no" "yes" "no" ...
 $ length    : num [1:8] 63 54 64 45 39 60 43 49
 $ distance  : num [1:8] 86 85 160 74 95 118 95 112
 - attr(*, "spec")=
  .. cols(
  ..   order = col_double(),
  ..   supplement = col_character(),
  ..   length = col_double(),
  ..   distance = col_double()
  .. )
```

The dataframe comprises 8 observations of 4 variables. Each row in the dataframe represents a record for an individual male bitterling. The variables are the numerical variable `order` which represents the order in which the males were tested, the categorical variable `supplement` with two levels; `no` and `yes`, indicating which individuals received a food supplement. The two other variables in the dataframe are `length` and `distance`, corresponding with individual male standard length (mm) and male response distance to an intruder (cm). These are both numerical continuous variables.

3.2.4.2 Frequentist linear model

We will proceed by fitting a simple (frequentist) general linear model (GLM) to obtain parameter estimates to use as priors. The model is fitted as:

```
p1 <- lm(distance ~ length + supplement, data = pilot)
```

A neat numerical output is obtained with the `tidy` function from the `broom` package:

```
broom::tidy(p1)
```

```
# A tibble: 3 x 5
  term          estimate std.error statistic p.value
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)      19.9     36.1       0.552  0.605
2 length            1.27     0.684     1.86   0.123
3 supplementyes    34.0     12.2       2.78   0.0390
```

1. The response distance of males when standard length is zero is approximately 20 cm (sd ~ 40). This is the `intercept`.
2. A 1 mm increase in male standard length results in an increased response distance of approximately 1.3 cm (sd ~ 0.7). This is the slope of `length`.
3. Supplementary feeding adjusted the slope (response distance) by approximately 35 cm (sd ~ 15) (`supplementyes`).

Since we did not include an interaction in the pilot model it will be incorporated as a weakly-informative prior in the full Bayesian model.

3.2.4.3 Priors on the fixed effects

These findings can be specified in the model as priors on the fixed effects as:

$\beta intercept \sim N(20, 1600)$ (mean, variance)

$\beta sl \sim N(1.3, 0.49)$

$\beta fSupp \sim N(35, 225)$

$\beta interaction \sim N(0, 1000)$

Thus, in the case of $\beta intercept$, we assume normality with a mean of 20 cm and variance of 1600 (sd = 40) cm.

3.2.4.4 Priors on the hyperparameter

The prior distribution on the hyperparameter should also be specified. The default is a diffuse gamma distribution, but other distributions can be used, for a full list of available distributions in INLA see:

```
names(inla.models()$prior)
```

In addition to these available prior distributions, it is also possible to define your own. In this model we will use a Gaussian distribution with a weakly-informative prior.

$\sigma \sim N(0, 1)$

Model variance is assumed to be normal, with a mean of 0 and variance of 1.

3.2.5 Fit the model

We will fit two Bayesian Gaussian GLMs using INLA, one with default priors (`M0`) and the second with informative priors on the fixed effects, derived from the pilot study, and weakly informative priors on the hyperparameter (`M1`).

The default INLA model is fitted with the following script:

```
M0 <- inla(resp_dist ~ sl * fSupp, data = bitt)
```

The default priors used for the model can be obtained with:

```
inla.priors.used(M0)
```

This output shows that for the fixed effects the priors for the default model are:

$\beta intercept \sim N(0, 0)$ $(\tau = 0)$

$\beta sl \sim N(0, 1000)$ $(\tau = 0.001)$

$\beta fSupp \sim N(0, 1000)$ $(\tau = 0.001)$

$\beta interaction \sim N(0, 1000)$ $(\tau = 0.001)$

And for the hyperparameter:

$\sigma \sim$ loggamma $(1, 2 \text{ x } 10^5)$ $(\tau = 1 \text{ x } 5^{-6})$

The model with informative priors is fitted with the following script:

```
M1 <- inla(resp_dist ~ sl * fSupp, data = bitt, control.family
= list(hyper = list(prec = list(prior = "gaussian", param
= c(0, 1)))), control.fixed = list(mean.intercept = 20,
prec.intercept = 40^(-2), mean = list(sl = 1.3, fSupp1
= 35, default = 0), prec = list(sl = 0.7^(-2), fSupp1 =
15^(-2), default = 31.62^(-2))))
```

The priors can be obtained with:

```
inla.priors.used(M1)
```

This output shows that for the fixed effects:

$\beta intercept$ ~ $N(20, 1600)$ (τ = 6.25 x 10^{-5})

βsl ~ $N(1.3, 0.49)$ (τ = 2.04)

$\beta fSupp$ ~ $N(35, 225)$ (τ = 4.44 x 10^{-3})

$\beta interaction$ ~ $N(0, 1000)$ (τ = 0.001)

And for the hyperparameter:

σ ~ $N(0, 1)$ (τ = 1)

3.2.6 Obtain the posterior distribution

3.2.6.1 Model with default priors

3.2.6.1.1 Fixed effects Output from model `M0` can be obtained with:

```
summary(M0)
```

However, this command produces an intimidating cascade of information (not shown here).

An alternative is to look first at the posterior mean, standard deviation and 95% credible intervals for the fixed effects:

```
M0Betas <- M0$summary.fixed[,c("mean", "sd", "0.025quant",
"0.975quant")]
```

```
round(M0Betas, digits = 2)
```

```
             mean    sd 0.025quant 0.975quant
(Intercept) 59.07 17.01      25.91      92.88
sl           1.77  0.32       1.13       2.40
fSupp1      32.35 21.50     -10.41      74.11
sl:fSupp1   -0.07  0.42      -0.88       0.75
```

This output reports the *posterior mean* and *standard deviation* for the model intercept (`intercept`), covariates (`sl`, `fSupp1`) and interaction (`sl:fSupp1`). Note that there are no P-values, which are used in frequentist analyses but are meaningless in a Bayesian context. Instead we have the 95% *credible intervals*; these are the 0.025 and 0.975 quantiles in the output above.

For the variable `sl` we have a posterior mean of 1.77 and lower 95% credible interval of 1.13 and upper 95% credible interval of 2.4. We can conclude from this result that we are 95% certain that the posterior mean of the regression parameter for `sl` falls between these credible intervals.

Because the credible intervals for `sl` do not encompass zero, we can be confident that the slope of the relationship is greater than zero. That is, we are 95% certain that the true value of the `sl` parameter in our model is between 1.13 and 2.4 given the data and (default) prior information provided to the model. In a Bayesian context we cannot consider this result 'significant,' because significance testing only applies in a frequentist hypothesis testing setting. However, we can conclude that `sl` is *statistically important* in the default model.

Similarly, we can conclude that the `Intercept` of the relationship, with credible intervals from 25.91 to 92.88, differs from zero with a posterior mean of 59.07 and standard deviation of 17.01.

For supplementary feeding (`fSupp1`), and the interaction between standard length and supplementary feeding (`sl:fSupp1`), the credible intervals range from negative values for the lower credible interval to positive for the upper interval, indicating that these model parameters do not differ from zero.

Instead of just summarising the posterior distribution of the fixed effects with a posterior mean and a 95% credible interval, we can plot the posterior distribution of each parameter, available in the INLA object `M0$marginals.fixed`. The posterior distributions can be visualized using `ggplot2`. The coding for this plot is available in the R script associated with this chapter.

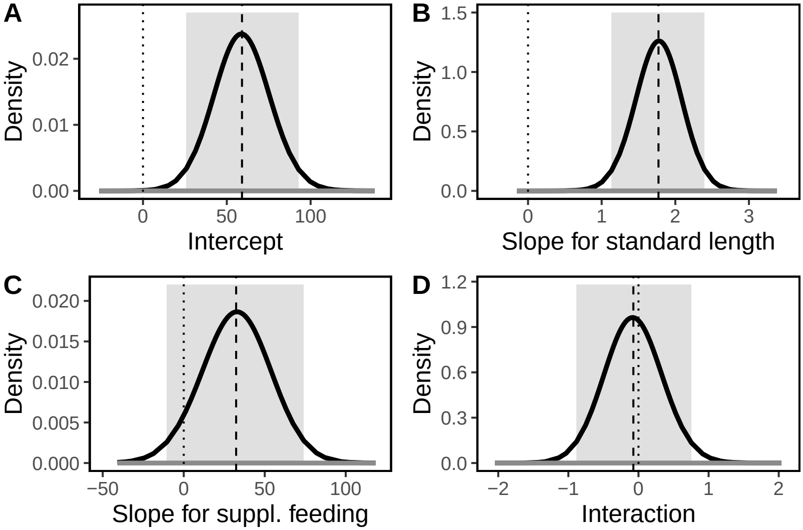

Figure 3.5: **Posterior and prior distributions for fixed parameters of a Bayesian linear regression to predict the territorial response distance of male European bitterling in response to a rival. The model is fitted with default (non-informative) priors. Distributions for: A. model intercept; B. slope for male standard length; C. slope for supplementary feeding; D. interaction of male standard length and supplementary feeding. The solid black line is the posterior distribution, the solid gray line is the prior distribution, the gray shaded area encompasses the 95% credible intervals, the vertical dashed line is the posterior mean of the parameter, the vertical dotted line indicates zero.**

Figure 3.5 provides a visual representation of the summary of the fixed effects. For parameters where zero (indicated by the dotted line) falls outside the range of the 95% credible intervals (gray shaded area), the parameter is considered statistically important. Thus, the intercept (panel A) and slope for male standard length (panel B) differ from zero and are statistically important, while the slope for supplementary feeding and interaction between standard length and supplementary feeding are not (i.e. panels C and D). This figure also shows the non-informative priors, which appear flat across the range of possible values (hence non-informative priors are sometimes termed 'flat' priors), and make a limited contribution to the posterior distribution.

3.2.6.1.2 Hyperparameter Model `M0` contains a parameter, sigma (σ), that is used for the variance (σ^2) of the normal distribution for male response distance. In the context of an INLA model, the variance parameter is termed a 'hyperparameter.' In a simple linear model like `M0` the hyperparameter just comprises the model residual variance.

As with the fixed effects, we can put priors on the hyperparameter (or use the non-informative default) but a vital step in fitting a Bayesian model is to examine the posterior distribution of the hyperparameter(s).

Recall that a slight complication is that INLA uses precision (τ or tau) rather than the variance of the hyperparameter, though this is simply the reciprocal of the variance.

We can obtain a summary of the precision of the hyperparameter with:

```
M0hyp <- M0$summary.hyper[,c("mean", "mode", "0.025quant",
"0.975quant")]
```

	mean	mode	0.025quant	0.975quant
Precision for Gaussian obs	0.004	0.0038	0.0025	0.0057

The posterior distribution of the precision of the hyperparameter can be visualized using `ggplot2`. See the R script associated with this chapter. Because the posterior distribution is not symmetrical, we plot the posterior mode (rather than mean) as a dashed vertical line.

Since we typically do not work with precision, we obtain the posterior distribution of the standard deviation of the hyperparameter (sigma, σ) with:

```
round(bri.hyperpar.summary(M0),2)
```

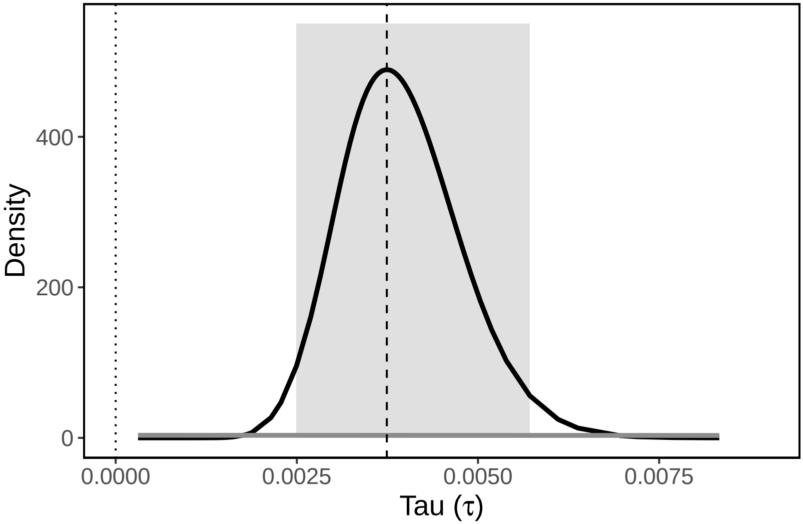

Figure 3.6: **Posterior and prior distributions for the precision of the hyperparameter of a Bayesian linear regression to predict the territorial response distance of male European bitterling to a rival. The model is fitted with default (non-informative) priors. The solid black line is the posterior distribution, the solid gray line is the prior distribution, the gray shaded area encompasses the 95% credible intervals, the vertical dashed line is the posterior mode, the vertical dotted line indicates zero.**

	mean	mode	0.025quant	0.975quant
SD for Gaussian observations	16.17	15.69	13.24	19.97

Visualisation of the posterior distribution of the standard deviation of the hyperparameter can be achieved with `ggplot2` using R script associated with this chapter.

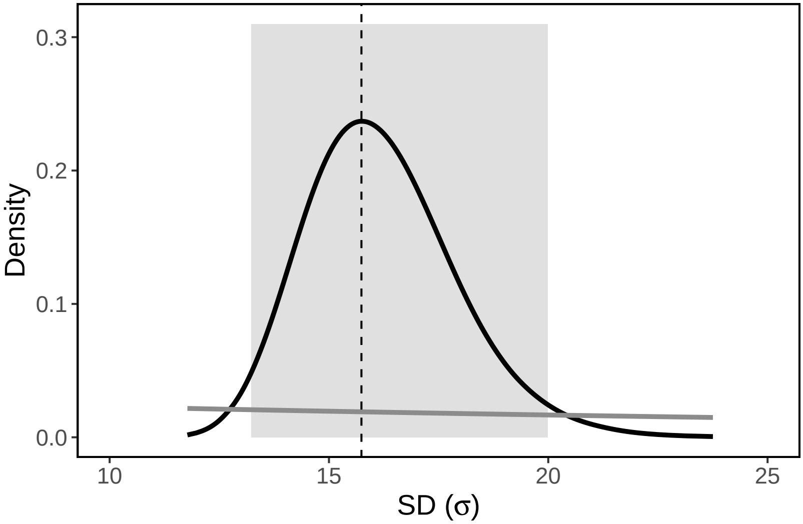

Figure 3.7: **Posterior and prior distributions for the standard deviation of the hyperparameter of a Bayesian linear regression to predict the territorial response distance of male European bitterling to a rival. The model is fitted with default (non-informative) priors. The solid black line is the posterior distribution, the solid gray line is the prior distribution, the gray shaded area encompasses the 95% credible intervals, the vertical dashed line is the posterior mode.**

Clearly the standard deviation of the hyperparameter differs from zero (Fig. 3.7). The distribution is also not normal; the default prior is for a gamma distribution.

3.2.6.2 Model with informative priors

As for the model with default priors, we will examine the posterior distributions for the model with informative priors, starting with the fixed effects.

3.2.6.2.1 Fixed effects First examine the posterior mean and 95% credible intervals for the fixed effects:

```
M1Betas <- M1$summary.fixed[,c("mean", "sd", "0.025quant",
"0.975quant")]
```

```
round(M1Betas, digits = 2)
```

```
             mean    sd 0.025quant 0.975quant
(Intercept) 55.09 12.56      30.42      79.81
sl           1.84  0.24       1.37       2.31
fSupp1      41.12 12.82      15.90      66.22
sl:fSupp1   -0.24  0.26      -0.74       0.27
```

This reports the posterior mean, standard deviation and 95% credible intervals for the `intercept`, covariates (`sl`, `fSupp1`) and interaction (`sl:fSupp1`). Note that the posterior means differ quantitatively from the default model as do the 95% credible intervals, which encompass a narrower range in each case.

For the variable `sl` we now have a posterior mean of the slope of 1.84 and lower 95% credible interval of 1.37 and upper 95% credible interval of 2.31. We can conclude from this result that we are 95% certain that the posterior mean of the regression parameter for the slope of sl falls between these credible intervals, centered around the posterior mean.

We can similarly conclude that the `Intercept` of the relationship differs from zero, with a posterior mean of 55.09 and credible intervals from 30.42 to 79.81.

For supplementary feeding (`fSupp1`), in contrast to the model with non-informative priors, the parameter is statistically important, with a posterior mean of 41.12 and 95% credible intervals from 15.9 to 66.22.

In the case of the interaction between standard length and supplementary feeding (`sl:fSupp1`) the credible intervals range from negative values for the lower interval (-0.74) to positive for the upper interval (0.27), indicating that this model parameter does not differ from zero.

The posterior distributions of the fixed effects can be visualized using ggplot2. The coding for this plot is available in the R script associated with this chapter.

Figure 3.8 indicates that for model M1 the intercept, slope for male standard length and slope for supplementary feeding all differ from zero and are statistically important in the model. The interaction between standard length and supplementary feeding is not. This figure also shows the distributions of the informative priors, based on the pilot study described in Section 3.2.4.1. These informative priors influence the posterior distribution.

3.2.6.2.2 Hyperparameter A summary of the precision of the hyperparameter for the informative model is obtained with:

```
M1hyp <- M1$summary.hyper[,c("mean", "mode", "0.025quant",
"0.975quant")]
```

	mean	mode	0.025quant	0.975quant
Precision for Gaussian obs	0.0048	0.0046	0.0032	0.0066

The posterior distribution of the precision of the hyperparameter can be visualized using ggplot2. The coding for this plot is available in the R script associated with this chapter.

While informative priors were put on fixed effects in the model, a weakly informative prior was put on the hyperparameter; evident in the prior distribution in Fig. 3.9. The 95% credible intervals of the posterior distribution of the hyperparameter do not include zero.

We can obtain the posterior distribution of the standard deviation of the hyperparameter (sigma, σ) with:

```
round(bri.hyperpar.summary(M1),2)
```

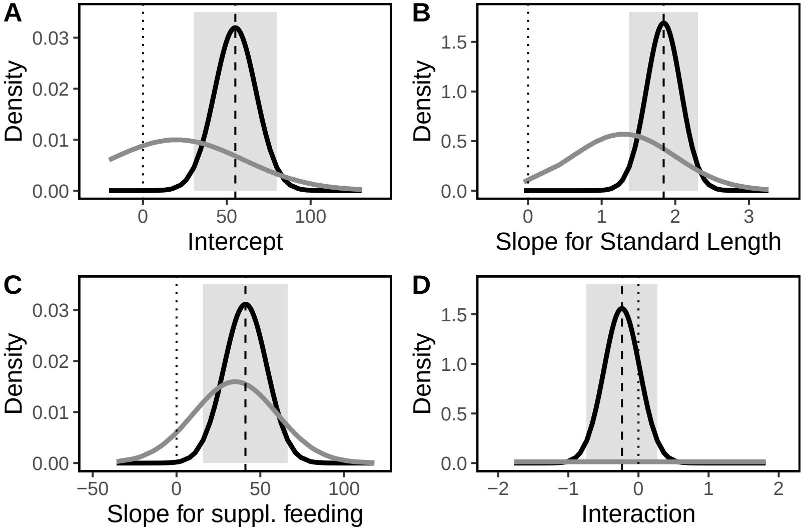

Figure 3.8: **Posterior and prior distributions for fixed parameters of a Bayesian linear regression to predict the territorial response distance of male European bitterling (*Rhodeus amarus*) in response to a rival fitted with informative priors. Distributions for: A. model intercept; B. slope for male standard length; C. slope for supplementary feeding; D. interaction of male standard length and supplementary feeding. The solid black line is the posterior distribution, the solid gray line is the prior distribution, the gray shaded area encompasses the 95% credible intervals, the vertical dashed line is the posterior mean of the parameter, the vertical dotted line indicates zero. For parameters where zero (indicated by dotted line) falls outside the range of the 95% credible intervals (gray shaded area), the parameter is considered statistically important (i.e. in the case of panels A, B and C).**

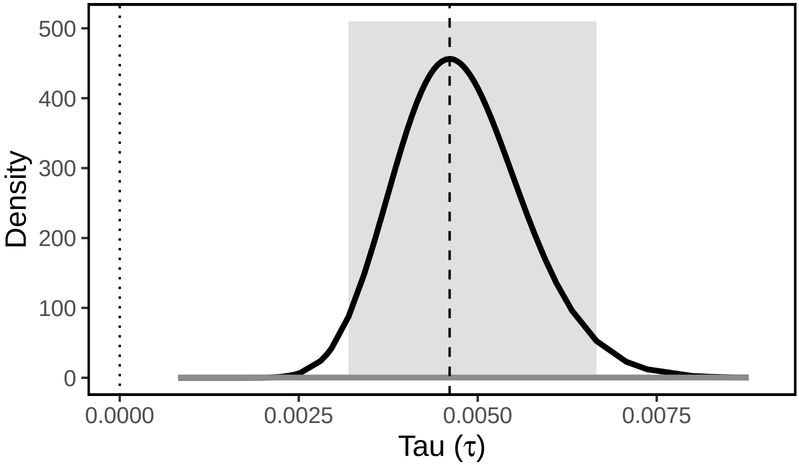

Figure 3.9: **Posterior distribution for the precision of the hyperparameter of a Bayesian linear regression to predict the territorial response distance of male European bitterling to a rival. The model is fitted with a weakly informative prior on the hyperparameter. The solid black line is the posterior distribution, solid gray line is the prior distribution, the gray shaded area encompasses the 95% credible intervals, the vertical dashed line is the posterior mode, the vertical dotted line indicates zero.**

	mean	mode	0.025quant	0.975quant
SD for Gaussian observations	14.67	14.30	14.56	17.67

Visualisation of the posterior distribution of the standard deviation of the hyperparameter can be accomplished with `ggplot2` using R script associated with this chapter.

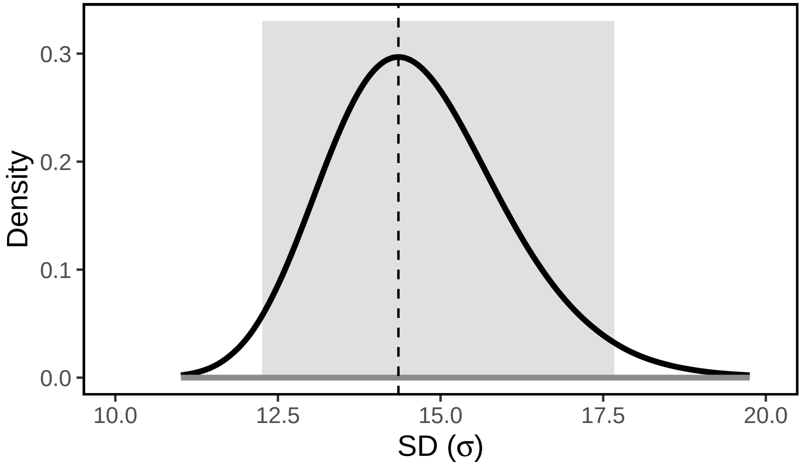

Figure 3.10: **Posterior and prior distributions for the standard deviation of the hyperparameter of a Bayesian linear regression to predict the territorial response distance of male European bitterling to a rival. The model is fitted with a weakly informative prior on the hyperparameter. The solid black line is the posterior distribution, the solid gray line is the prior distribution, the gray shaded area encompasses the 95% credible intervals, the vertical dashed line is the posterior mode.**

3.2.6.3 Comparison with frequentist Gaussian GLM

At this stage it is instructive to compare the results of the Bayesian Gaussian GLMs with the same model fitted in a frequentist setting. Execution of the model in a frequentist framework can be performed with:

```
Freq <- lm(resp_dist ~ sl * fSupp, data = bitt)
```

The results are obtained with:

```
broom::tidy(Freq)%>% mutate_if(is.numeric, round, 4)
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)   47.7      18.8        2.54  0.0148
2 sl             1.98      0.353      5.61  0
3 fSupp1        59.1      28.9        2.05  0.0467
4 sl:fSupp1     -0.580     0.554     -1.05  0.301
```

We already have the results for the Bayesian models; for the model with default priors these are:

```
round(M0Betas, digits = 2)
```

```
             mean    sd 0.025quant 0.975quant
(Intercept) 59.07 17.01      25.91      92.88
sl           1.77  0.32       1.13       2.40
fSupp1      32.35 21.50     -10.41      74.11
sl:fSupp1   -0.07  0.42      -0.88       0.75
```

For the Bayesian model with informative priors:

```
round(M1Betas, digits = 2)
```

```
             mean    sd 0.025quant 0.975quant
(Intercept) 55.09 12.56      30.42      79.81
sl           1.84  0.24       1.37       2.31
fSupp1      41.12 12.82      15.90      66.22
sl:fSupp1   -0.24  0.26      -0.74       0.27
```

These results can be summarised together in a table:

Table 4.1: **Parameters for fixed effects of a model to investigate the effect of standard length (sl) and supplementary feeding treatment (fSupp) and their interaction on the territorial response distance of male European bitterling for a frequentist**

GLM, Bayesian GLM with default priors and Bayesian GLM with informative priors. Mean (sd) parameter estimates are shown for each model

Model	Intercept	sl	fSupp	sl : fSupp1
Frequentist	47.7(18.8)	2.0(0.4)	59.1(28.9)	-0.6(0.6)
Bayesian (default)	59.1(17.0)	1.8(0.3)	32.4(21.5)	-0.1(0.4)
Bayesian (informative)	55.3(12.5)	1.8(0.2)	40.5(12.4)	-0.2(0.3)

While parameter estimates for the frequentist and Bayesian models are broadly similar, it is notable that results for the Bayesian model with default (non-informative) priors diverge more from the results for the frequentist model than do the parameter estimates for the Bayesian model with informative priors.

It is a common misconception that non-informative priors are objective and provide an unbiased representation of the data. However, 'non-informative' is a misnomer, because all priors influence model outcomes. In a Bayesian framework, the implementation of carefully specified informative priors will typically be more likely to generate robust results than reliance on default priors.

We can also compare the standard deviation of the residuals (sigmas) for these models.

For the Frequentist model:

```
round(summary(Freq)$sigma,2)
```

```
16.22
```

For the Bayesian model with default priors:

```
round(bri.hyperpar.summary(M0)[,c("mean")],2)
```

```
16.19
```

For the Bayesian model with informative priors:

```
round(bri.hyperpar.summary(M1)[,c("mean")],2)
```

```
14.67
```

Estimates of sigma are almost identical for the frequentist and Bayesian model with default priors. The greater precision of the Bayesian model with informative priors is reflected by a smaller sigma.

3.2.7 Conduct model checks

After model fitting and obtaining the posterior distributions, an important next step is validation of the model through model checks. At this stage we may also wish to perform model selection.

3.2.7.1 Model selection using the Deviance Information Criterion (DIC)

When a model is fitted with several explanatory variables, including interaction terms, we have the opportunity to conduct *model selection.* Model selection involves finding an optimal set of covariates for a model. It is a hotly debated subject in statistics, with several alternative approaches. Here we present a simple model selection procedure for models `M0` and `M1`. A more sophisticated model selection procedure using an Information Theoretic (IT) approach is presented in Chapter 5.

In a frequentist setting a common approach to model selection is to use classical backward or forward stepwise model selection based on the Akaike Information Criteria (AIC). AIC measures goodness-of-fit and model complexity, with the lower the AIC score, the better the fit of the model to the data, penalised by model complexity. In backward model selection, a model with all covariates is fitted and then sequential deletion of covariates is undertaken until removal of further covariates fails to improve the fit of the model. In forward selection this procedure is reversed.

In a Bayesian framework the Deviance Information Criterion (DIC) can similarly be used to compare model goodness-of-fit while penalising model complexity. Like AIC, a smaller DIC score indicates a better fit of the model to the data given its complexity.

A model's DIC score can be computed in INLA using the `dic = TRUE` option in `control.compute`. For model `M0`:

```
M0 <- inla(resp_dist ~ sl * fSupp, control.compute =
list(dic = TRUE), data = bitt)
```

And the same can be computed for `M1`.

For `INLA` there is no stepwise model selection procedure (such as the `drop1` and `step` functions for frequentist GLMs), which means model selection must be conducted manually.

The goal in conducting model selection in this case is twofold:

1. Compare full and reduced models for models with non-informative and informative priors.
2. Compare best-fitting models with non-informative and informative priors.

Start by sequentially removing model parameters from `M0` and then compare using the DIC:

The full model:

```
M0.full <- inla(resp_dist ~ sl * fSupp, control.compute =
list(dic = TRUE), data = bitt)
```

Drop interaction:

```
M0.1 <- inla(resp_dist ~ sl + fSupp, control.compute =
list(dic = TRUE), data = bitt)
```

Drop supplementary feeding:

```
M0.2 <- inla(resp_dist ~ sl, control.compute = list(dic =
TRUE), data = bitt)
```

Drop standard length:

```
M0.3 <- inla(resp_dist ~ fSupp, control.compute = list(dic
= TRUE), data = bitt)
```

Compare with the DIC:

```
DIC <- cbind(c(M0.full$dic$dic, M0.1$dic$dic, M0.2$dic$dic,
M0.3$dic$dic))    rownames(DIC) <- c("full","no inter","no
suppl","no sl") round(DIC,1)
```

```
           DIC
full     409.7
no inter 408.9
no suppl 436.6
no sl    438.0
```

The model without an interaction generates the lowest DIC score (408.9). This score is only marginally lower than the score for the full model with the interaction included, which is an indication that the interaction is not important in the model. A difference in DIC scores of between 5 and 10 would be considered substantial.

Following the same procedure with model M1 (see R script associated with this chapter) yields:

```
            DIC
full      408.4
no inter  408.5
no suppl  436.7
no sl     438.2
```

In this case the full model, with an interaction, generates the lowest DIC score (408.4). However, as in the case above, this score is only marginally lower than the score for the model without the interaction, which tells us that the interaction is not important.

We can conclude that both for the model with non-informative and informative priors that the best-fitting model in each case is the one that includes both `sl` and `fSupp`, but with no interaction between them.

We can now compare the best-fitting models with non-informative and informative priors using the DIC:

```
DIC2 <- cbind(c(M0.1$dic$dic, M1.1$dic$dic))

rownames(DIC2) <- c("default priors","informative priors")

colnames(DIC2) <- "DIC"

round(DIC2,2)
```

```
                      DIC
default priors     408.92
informative priors 408.47
```

These DIC score are essentially the same.

Given the similarity in goodness-of-fit of both these models, what should we do? Since the DIC scores for both models are essentially the same, the appropriate course is to continue with model checking for both and present the findings for both models. For brevity, however, we will continue by examining the model with informative priors only.

3.2.7.2 Posterior predictive checks

The purpose of posterior predictive checks is to assess if a model generates realistic predictions. It does this by drawing simulated estimates from the joint posterior predictive distribution and comparing them with observed data. Any departure of the simulated data from the observed data will reflect problems with the model. Ideally, simulated data should match the observed. This matching is performed with a posterior predictive p-value. If the posterior predictive p-value is close to 0.5 it means simulated and observed data are similar. However, if a posterior predictive p-value is close to 1 it means the model prediction is too high, if close to zero, the model prediction is too low. A frequency plot of posterior predictive p-values should show a distribution centred around 0.5.

In `INLA` the posterior predictive p-value can be obtained with the function `inla.pmarginal()`. See the R script associated with this chapter for estimating and plotting the posterior predictive p-values for the Bayesian model with informative priors without interaction.

The frequency histogram of posterior predictive p-values in Fig. 3.11 shows that most values are close to zero or 1, with few close to 0.5, which indicates the model check has not been satisfied; the data are overdispersed compared to the model. Consequently, we will proceed with further model checks.

3.2.7.3 Cross-validation model checking

Cross validation is a model-checking approach that examines how well a model is able to generalise to new data. Leave-one-out cross validation (LOO-CV) involves systematically dropping a single data point, refitting the model and evaluating the altered model inference. Following (Wang et al., 2018), we use the *conditional predictive ordinate* (CPO) and *probability integral transform* (PIT) to evaluate model goodness-of-fit. To obtain both we simply run the model using the `cpo = TRUE` option in `control.compute`.

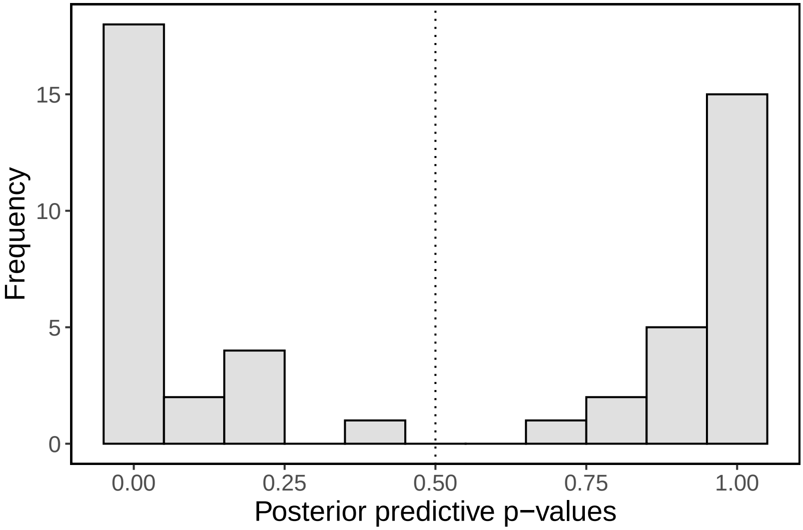

Figure 3.11: **Frequency histogram of the posterior predictive p-values for the best-fitting Bayesian linear regression with informative priors to predict the territorial response distance of male European bitterling to a rival. The vertical dotted line indicates 0.5.**

To ensure there are no potential numerical problems in estimating CPO or PIT for a given model, we first run the following check:

```
sum(M1.pred$cpo$failure)
```

0

An outcome of zero indicates no problems with the computation of CPO or PIT. A value of 1 would indicate CPO or PIT were unreliable.

Plotting PIT values will indicate whether the predictive distributions match the data, apparent as a uniform distribution. We can assess uniformity visually with a frequency histogram and Q-Q plot of PIT values for a uniform distribution (see the R script associated with this chapter).

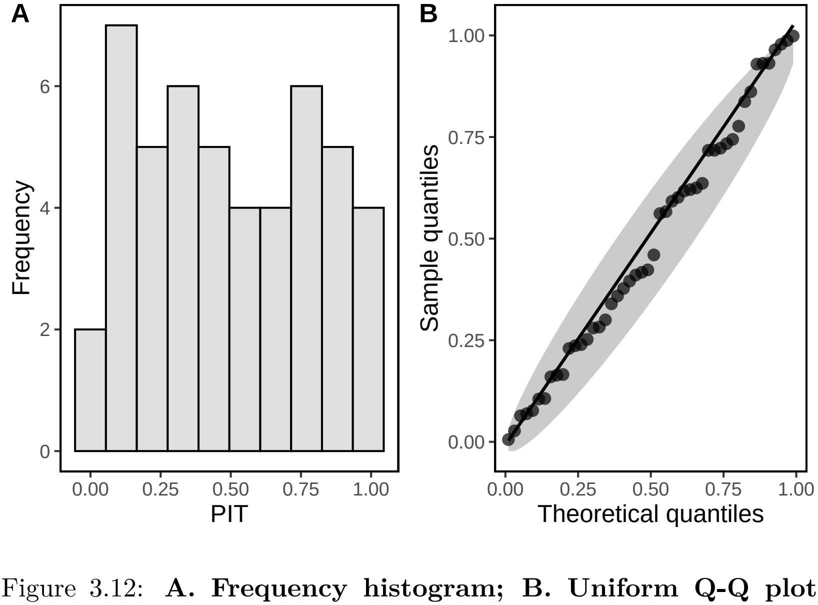

Figure 3.12: **A. Frequency histogram; B. Uniform Q-Q plot with confidence bands (shaded gray), for cross-validated probability integral transform (PIT) values for the best-fitting Bayesian linear regression with informative priors.**

The frequency histogram of PIT values in Fig. 3.12 A shows that the distribution is broadly uniform, with no clustering of values at zero or 1. This conclusion is supported by the Q-Q plot (Fig. 3.12 B), which shows that the PIT values match a uniform distribution.

3.2.7.4 Bayesian residuals analysis

Homogeneity of residual variance can be assessed visually by plotting model residual variance against fitted values as well as each variable in the model (see the R script associated with this chapter).

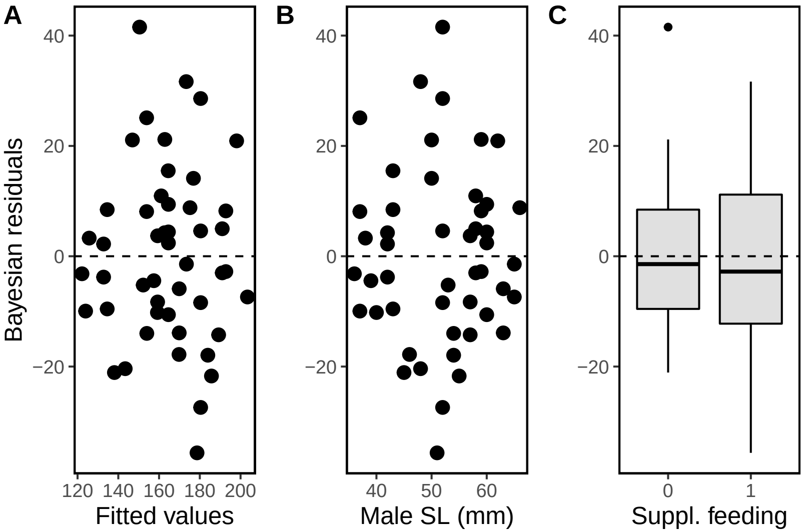

Figure 3.13: **Bayesian residuals plotted against: A. fitted values ; B. male standard length; and C. supplementary feeding, to assess homogeneity of residual variance.**

Ideally, the distribution of residuals around zero should be random along the horizontal axis, which is the case in Fig. 3.13 A and B, and in the case of a categorical variable, the median of a boxplot of residual values should be approximately zero, which is the case in Fig. 3.13 C.

3.2.7.5 Prior sensitivity analysis

A final Bayesian model check is to examine prior distributions through a sensitivity analysis. This procedure is important both in the case of non-informative and informative priors. The procedure involves systematically changing prior distributions and examining the magnitude of outcome for the posterior distribution.

To investigate the impact of different priors, we increased and decreased priors on the fixed effects by 20% and examined the outcome for the posterior mean.

The original priors for the fixed effects were:

$\beta intercept \sim N(20, 1600)$

$\beta sl \sim N(1.3, 0.49)$

$\beta fSupp1 \sim N(35, 225)$

In the case of an increase by 20%, the priors for the fixed effects are:

$\beta intercept \sim N(24, 1920)$

$\beta sl \sim N(1.56, 0.59)$

$\beta fSupp1 \sim N(42, 270)$

In the case of a decrease by 20%, the priors for the fixed effects are:

$\beta intercept \sim N(16, 1280)$

$\beta sl \sim N(1.04, 0.39)$

$\beta fSupp1 \sim N(28, 180)$

Two alternative models were fitted with these increases and decreases in the priors and estimates for the betas obtained (see the R script associated with this chapter).

Table 4.2: **Sensitivity analysis for a 20% increase and decrease in priors on fixed effects and the % change in the posterior mean.**

Parameter	% prior	Mean	0.025CI	0.975CI	% posterior
	+20	57.6	33.5	81.5	-1.68
Intercept	0	58.6	34.9	82.1	0
	-20	60.0	37.0	83.0	2.44
	+20	1.8	1.3	2.2	0.92
sl	0	1.8	1.3	2.2	0
	-20	1.7	1.3	2.2	-1.35
	+20	30.3	22.2	38.5	1.18
fSupp1	0	30.0	21.9	38.1	0

Parameter	% prior	Mean	0.025CI	0.975CI	% posterior
	-20	29.4	21.4	37.5	-1.69

The results of the prior sensitivity analysis show that changes as large as 20% (increase and decrease) result in negligible changes to the posterior distribution.

We can plot the posterior distributions of these alternative models to visualise the changes (see R script associated with this chapter).

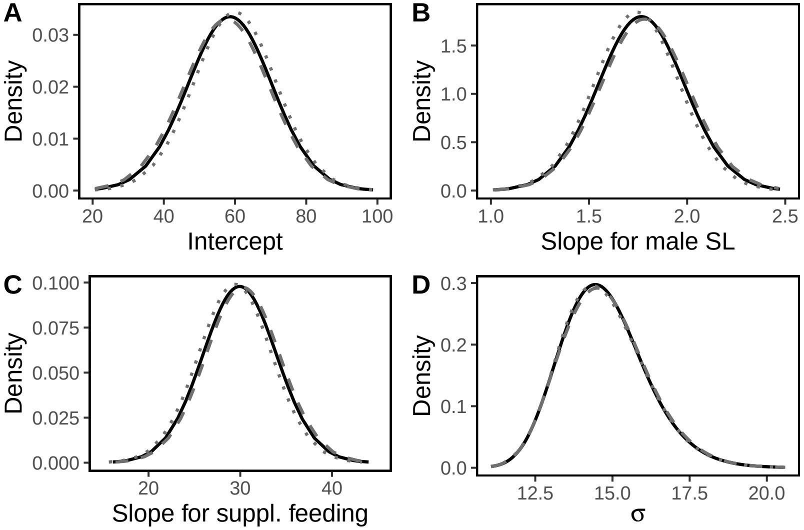

Figure 3.14: **Posterior distributions for parameters of a Bayesian linear regression to predict the territorial response distance of male European bitterling in response to a rival. Distributions for: A. model intercept; B. slope for male standard length; C. slope for supplementary feeding; D. hyperparameter. The solid black line is the posterior distribution for the optimal model, the dashed gray line is the posterior distribution for an alternative model with the priors increased by 20%, the dotted gray line is the posterior distribution for an alternative model with the priors decreased by 20%.**

Plots of the posterior distributions for the fixed effects and hyperparameter (Fig. 3.14) further illustrate that the posterior distributions of model parameters are robust to changes in the priors.

3.2.7.6 Conclusions from model checks

Manual model selection based on the DIC allowed us to slightly refine the model by dropping the interaction between male standard length and the provision of supplementary food. The model with informative priors showed a comparable goodness-of-fit to that of the model with default priors. A plot of posterior predictive p-values suggested some overdispersion of the model, though leave-one-out cross validation indicated that the predictive distributions matched the data well. Residuals plots failed to highlight anything problematic with the model fit. Prior sensitivity analysis demonstrated the model to be robust to changes in prior distributions of fixed effects. Overall, then, the Bayesian GLM with informative priors appears to provide a good representation of the data.

3.2.8 Interpret and present model output

We specify the Bayesian GLM using mathematical notation in exactly the same way as we would for a frequentist model:

$Response_i \sim Gaussian(\mu_i, \sigma^2)$

$E(Response_i) = \mu_i$ and $\text{var}(Response_i) = \sigma^2$

$\mu_i = \beta_1 + \beta_2 \text{ x } Length_i + \beta_3 \text{ x } Supplement_i$

Where $Response_i$ is the aggressive response distance (cm) of male European bitterling i assuming a normal distribution with mean μ_i and variance σ^2. $Length_i$ is a continuous covariate representing the standard length of male bitterling i (mm) and $Supplement_i$ is a categorical variable representing the provision of supplementary food to male i, with two levels; supplement provided or no supplement. The numerical output for the fixed effects of the final model is:

```
             mean    sd 0.025quant 0.975quant
(Intercept) 58.56 12.01      34.88      82.11
```

```
sl           1.77  0.22        1.33       2.21
fSupp1      29.95  4.12       21.87      38.08
```

And for sigma:

```
      mean   q0.025   q0.975  term
1 14.73564 12.33454 17.73963 sigma
```

These results can be more formally presented in the following way:

Table 4.3: **Posterior mean estimates for aggressive response distances (cm) of male European bitterling (*Rhodeus amarus*) as a function of male standard length (mm) and a supplementary feeding treatment, modelled using a Gaussian GLM fitted using Bayesian inference with INLA. CrI are the Bayesian 95% credible intervals.**

Model parameter	Posterior mean	Lower 95% CrI	Upper 95% CrI
Intercept	58.56	34.92	82.09
Standard length	1.77	1.33	2.21
Supplementary feeding	29.25	21.88	38.07
σ	14.75	12.32	17.75

These results show a statistically important positive effect of male bitterling standard length on response distance, with larger males initiating attacks on a rival at greater distances than smaller males. The effect of supplementary feeding for 6-days prior to testing was similarly to increase the aggressive response distance to a rival.

3.2.9 Visualise the results

The final of the 9 steps to fitting a Bayesian GLM is to visualise the model (Section 3.2). A figure helps with understanding model outcomes and is a valuable summary of the model findings for a paper, thesis or report. The full coding for this plot is available in the R script associated with this chapter.

We start by defining a dataframe (‘`MyData`’) that contains `sl` and `fSupp` using `dplyr` functions:

```
MyData <- ddply(bitt, .(fSupp), summarize, sl = seq(from =
min(sl), to = max(sl), length = 50))
```

This creates 100 artificial covariate values. There is no `predict` function in INLA, but we can obtain fitted values manually with a design matrix for the values in `MyData` and then multiply this with the posterior mean values of the model.

We must also add an extra variable for the response variable to `MyData` and assign it 'NA.' We will then combine the `bitt` and `MyData` objects, and apply INLA to this combined data set. INLA will predict the response variable where an NA occurs.

```
MyData$resp_dist <- NA

bitt.Pred <- bitt[, colnames(MyData)]

bitt.Comb <- rbind(bitt.Pred, MyData)
```

We next re-run the model in `INLA` using the combined data set (`bitt.Comb`), ensuring that `compute = TRUE` is selected in the `control.predictor` argument:

```
Final.Pred <- inla(resp_dist ~ sl + fSupp,  data = bitt.Comb,
control.predictor = list(compute = TRUE), control.family
= list(hyper = list(prec = list(prior = "gaussian",param
= c(0,1)))), control.fixed = list(mean.intercept = 20,
prec.intercept = 40^(-2), mean = list(sl = 1.3, fSupp1
= 35, default = 0), prec = list(sl = 0.7^(-2), fSupp1 =
15^(-2), default = 1000)))
```

Generate predicted values and relevant components in `MyData`

```
Pred <- Final.Pred$summary.fitted.values[((nrow(bitt))+1):
(nrow(bitt) + nrow(MyData)),]

MyData$mu    <- Pred[,"mean"]

MyData$selow <- Pred[,"0.025quant"]

MyData$seup  <- Pred[,"0.975quant"]
```

Create figure labels:

```
label_supp <- c("0" = "No food supplement", "1" = "With
food supplement")
```

And plot with `ggplot2` (see the R script associated with this chapter).

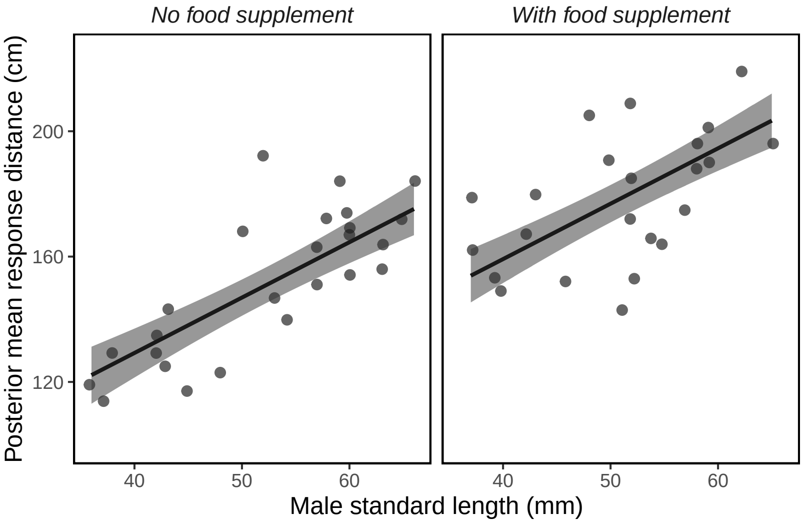

Figure 3.15: **Posterior mean aggressive response distance (cm) of male European bitterling (*Rhodeus amarus*) as a function of male standard length (mm) and supplementary feeding, modelled using a Gaussian GLM fitted using Bayesian inference with INLA. Shaded areas are Bayesian 95% credible intervals. Black points are observed data for different males.**

The results of this statistical analysis can be summarised as follows:

A Gaussian GLM was fitted to data using Bayesian inference with INLA to model the aggressive response distance (in cm) of a group of 48 territorial male European male bitterling (Rhodeus amarus) *to a model rival. There was a statistically important positive effect of male standard length (in mm) and supplementary feeding on response distance (Fig. 3.15). The mean slope of the relationship between response distance (cm) and standard length (mm) was 1.77 with 95% certainty that it lay between 1.33 and 2.21 (Table 4.3). The effect of supplementary feeding for six days prior to testing was to increase male response distance by 30 cm, with 95% certainty that it lay between 22 and 38 cm (Table 4.3). The model was fitted using informative priors on the fixed effects, obtained from a separate pilot study, and weakly informative effects on the hyperparameter.*

3.3 Conclusions

Bayesian inference offers an alternative approach to data analysis and has a number of advantages. One is that prior information can be incorporated into an analysis. Using prior information in a model is intuitively appealing and better reflects the scientific method of building on previous knowledge. A second advantage is in avoiding hypothesis testing and P-values, which do not allow us to draw direct conclusions about model parameters – only about hypothetical datasets (that we will never collect). Finally, there is a large range of advanced statistical methods that can only be performed in a Bayesian setting.

While a Bayesian model adds a layer of complexity to model fitting, since a careful consideration of the priors to be used is needed, it also adds an extra dimension to the sophistication of the analysis since, instead of simply presenting a model that describes the data, we now have a mechanism for incorporating previous knowledge or expert opinion through the prior distributions we put on model parameters.

Finally, the GLM fitted here using `INLA` demonstrates the user-friendliness of this package, as well as its flexibility, repeatability and computational speed in comparison with MCMC.

Chapter 4

Bayesian Poisson GLM

Placeholder

4.1 Stickleback lateral plate number

4.2 Steps in fitting a Bayesian GLM

4.2.1 State the question

4.2.2 Data exploration

4.2.2.1 Outliers

4.2.2.2 Distribution of the dependent variable

4.2.2.3 Balance of categorical variables

4.2.2.4 Multicollinearity among covariates

4.2.2.5 Zeros in the response variable

4.2.2.6 Relationships among dependent and independent variables

4.2.2.7 Independence of response variable

4.2.3 Selection of a statistical model

4.2.4 Specification of priors

4.2.4.1 Existing data

4.2.4.2 Priors on the fixed effects

4.2.5 Fit the models

4.2.6 Obtain the posterior distribution

4.2.6.1 Model with default priors

4.2.6.2 Model with informative priors

4.2.6.3 Fixed effects

4.2.6.4 Comparison with frequentist Poisson GLM

4.2.7 Conduct model checks

4.2.7.2.2 Simulate regression parameters from the posterior distribution

4.2.7.2.3 Calculate predicted values

4.2.7.2.4 Simulate count data using `rpois`

4.2.7.2.5 Calculate summary statistic

4.2.7.2.6 Repeat simulation

4.2.7.2.7 Compare dispersion of simulated and observed data

4.2.7.3 Posterior predictive checks

4.2.7.4 Cross-validation model checking

4.2.7.5 Bayesian residuals analysis

4.2.7.6 Prior sensitivity analysis

4.2.7.7 Conclusions from model checks

4.2.8 Interpret and present model output

4.2.9 Visualise the results

4.3 Conclusions

Chapter 5

Bayesian negative binomial GLM

Placeholder

5.1 Coral abundance

5.2 Steps in fitting a Bayesian GLM

5.2.1 State the question

5.2.2 Data exploration

5.2.2.1 Outliers

5.2.2.2 Distribution of the dependent variable

5.2.2.3 Balance of categorical variables

5.2.2.4 Multicollinearity among covariates

5.2.2.5 Zeros in the response variable

5.2.2.6 Relationships among dependent and independent variables

5.2.2.7 Independence of response variable

5.2.3 Selection of a statistical model

5.2.4 Specification of priors

5.2.4.1 Previous study

5.2.4.2 Priors on the fixed effects

5.2.5 Fit the models

5.2.6 Obtain the posterior distribution

5.2.6.1 Model with default priors

5.2.6.2 Model with informative priors

5.2.6.3 Comparison of models with uninformative and informative priors

5.2.6.4 Comparison with frequentist Poisson GLM

Chapter 6

Bayesian Bernoulli GLM

Placeholder

6.1 Common cuckoo parasitism of great reed warbler nests

6.2 Steps in fitting a Bayesian GLM

6.2.1 State the question

6.2.2 Data exploration

6.2.2.1 Outliers

6.2.2.2 Distribution of the dependent variable

6.2.2.3 Balance of categorical variables

6.2.2.4 Multicollinearity among covariates

6.2.2.5 Zeros in the response variable

6.2.2.6 Relationships among dependent and independent variables

6.2.2.7 Independence of response variable

6.2.3 Selection of a statistical model

6.2.4 Specification of priors

6.2.4.1 Pilot study

6.2.4.2 Model pilot data

6.2.4.3 Priors on the fixed effects

6.2.5 Fit the models

6.2.6 Obtain the posterior distribution

6.2.6.1 Model with default priors

6.2.6.2 Model with informative priors

6.2.6.3 Comparison of models with uninformative and informative priors

Chapter 7

Bayesian gamma GLM

Placeholder

7.1 Common seal dive duration

7.2 Steps in fitting a Bayesian GLM

7.2.1 State the question

7.2.2 Data exploration

7.2.2.1 Outliers

7.2.2.2 Distribution of the dependent variable

7.2.2.3 Balance of categorical variables

7.2.2.4 Multicollinearity among covariates

7.2.2.5 Zeros in the response variable

7.2.2.6 Relationships among dependent and independent variables

7.2.2.7 Independence of response variable

7.2.3 Selection of a statistical model

7.2.4 Specification of priors

7.2.4.1 Priors on the fixed effects

7.2.4.2 Priors on the hyperparameter

7.2.5 Fit the model

7.2.6 Obtain the posterior distribution

7.2.6.1 Model with default priors

7.2.6.2 Model with informative priors

7.2.6.3 Comparison of models with uninformative and informative priors

7.2.6.4 Comparison with frequentist gamma GLM

Chapter 8

Implementing and assessing Bayesian GLMs

Placeholder

8.1 Prior information

8.1.1 The results of previous research

8.1.2 Logical considerations

8.1.3 Expert knowledge

8.1.4 Pilot data

8.2 Presenting the results of a Bayesian GLM

8.3 Reviewing Bayesian GLMs

8.4 Misuse of Bayesian inference

8.5 Conclusions

Chapter 9

Coda

How can we use new data to change what we currently believe? As ecologists we often make decisions in the face of uncertainty and incomplete information. Bayesian inference offers a framework for incrementally accruing scientific knowledge by explicitly building on the conclusions of previous findings

However, despite the attraction in using Bayesian inference to tackle ecological questions, there are many pitfalls to its implementation. Sovereign against many of these problems is transparency; clearly reporting how priors were obtained, why they are specified as they are, careful description of their impacts, and presentation of sensitivity analyses. Ultimately, Bayesian methods do not offer a panacea, but they are a valuable tool for the ecologist that encourages full use of the available data - whatever form those data take.

We hope this book is useful in extending your understanding of Bayesian data analysis with R. We are always interested to receive feedback; positive or negative, and also welcome questions about your own analyses; feel free to email us.

Wang, X., Yue, Y., & Faraway, J. J. (2018). *Bayesian regression modeling with inla*. Chapman; Hall/CRC. https://doi.org/10.1201/9781351165761

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2009). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-

210x.2009.00001.x