University of Newcastle upon Tyne

MAS8391: Project

Bayesian Analysis of Segment Number Variation in Centipedes

Author:
Charlotte Holland

Supervisor:
Dr. Malcolm Farrow

April 29, 2014

Abstract

In most groups of arthropods the number of segments for each species is fixed. Geophilomorph centipedes have been found to be an exception to this rule and present intraspecific variation in segment number, which is determined during embryonic development. It is difficult to say exactly what determines the number of segments developed although studies can be conducted to find factors that may have an influence. Bayesian techniques can be applied to draw meaningful conclusions from samples of data. Ordinal logistic regression models for segment numbers are used. There is strong evidence to suggest a latitude effect and a sex effect. Also, there is evidence to support a genetic effect and that there may be an environmental effect based on the conditions during embryonic development.

Acknowledgements

I would like to thank Dr. M. Farrow for his guidance, advise and time towards producing this project. I would also like to thank Prof. W. Arthur for providing the data and idea for the project.

Contents

Chapter 1

Introduction

The development of segment number variation in the arthropod species *Strigamia maritima* is something that has been of interest. Kettle and Arthur (2000) and Vedel et al. (2009) have conducted studies in which samples of centipedes have been taken to observe how different factors have an effect on their development. The studies have analysed the data using only frequentist methods. In this project, we wish to apply Bayesian analysis to see if we can improve on the current findings and extract further conclusions.

In our Bayesian analysis, we use Markov Chain Monte Carlo (MCMC) methods and utilise the computer software package `rjags`. Guidance and advice on how to use `rjags` can be found in Plummer (2014).

1.1 Content of the project

In Chapter 2 we introduce the ideas involved in Bayesian analysis and the type of modeling we will be using. Chapter 3 gives an insight into the arthropod species *S. maritima* and begins to introduce the experiments done by Kettle and Arthur (2000) that explore the effects of latitude on the segment number variation. We then create our model and begin to apply the Bayesian methods and analysis to the data in Chapter 4. To improve our analysis, in Chapter 5 we adapt our model to include a variance component to allow for a random effect at each latitude and then add a further variance component to allow for a random cluster effect in Chapter 6. Finally, in Chapter 7 we look at the studies conducted by Vedel et al. (2009) to observe a potential genetic effect occurring in the development in segment numbers in centipedes.

Chapter 2

Bayesian Inference and Markov Chain Monte Carlo Methods

2.1 Introduction to Bayesian inference

In Bayesian inference, uncertainty about unknown quantities is represented by a probability distribution. When data are observed the prior distribution is changed to give the posterior distribution using Bayes' theorem which shows $Posterior \propto Prior \times Likelihood$.

Many of the calculations involved in deriving important properties, such as the posterior means, variances, predictive distributions and marginal density functions, all use suitable integration methods. For many complex models, these integrals can prove to be difficult to evaluate, resulting in a need for other methods, namely stochastic simulation. We use Markov Chain Monte Carlo (MCMC) methods to perform our inference.

2.2 Markov Chain Monte Carlo

A Markov chain is a stochastic system in which a system moves from one state to another. It has the property that the next state we will be in depends only on the current state we are in and not on any previous state. Markov Chain Monte Carlo (MCMC) methods are a collection of algorithms that work by repeatedly drawing sample values of unknown quantities from a complex model in order to produce an approximation to the posterior distribution. The repeated successive simulations form a Markov chain whose stationary distribution is the desired distribution. Due to the construction of these simulations, successive observations are correlated. Brooks (1998) shows how to implement techniques used for MCMC.

Smith and Gelfand (1992) note that, using MCMC methods to take large samples from a density, we can approximately recreate the density using techniques such as kernel density estimates. As a result there is essentially a duality between the sample produced and the density from which it came. This allows scope for experimenting with prior specification in order to influence posterior outcomes.

2.3 The Gibbs sampler

The Gibbs sampler is a method that allows us to simulate samples from a normalized joint density function. These samples may be marginalised to obtain marginal distributions related to the joint density. Gelfand (2000) demonstrates the motivation for using Gibbs sampling methods for inference within a Bayesian framework, to remove the earlier integration difficulties. Specifically, the Gibbs sampler allows us to simulate random variables from a distribution indirectly, removing the need to calculate the density. It proves to be useful when it is difficult to write down the joint density function or integrate over it.

Gelfand and Smith (1990) describe the processes involved. For the set of variables $\theta_1, \theta_2, ..., \theta_k$, the outline of the algorithm is as follows

Step 1: Initialise the chain. Set arbitrary set of starting values to be $\theta_1^{(0)}\theta_2^{(0)}, ..., \theta_k^{(0)}$.
Set counter $j = 1$

Step 2: Draw:
$\theta_1^{(j)} \sim [\theta_1|\theta_2^{(j-1)}, ..., \theta_k^{(j-1)}]$
$\theta_2^{(j)} \sim [\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)} ..., \theta_k^{(j-1)}]$
$\vdots$
$\theta_k^{(j)} \sim [\theta_k|\theta_1^{(j)}, ..., \theta_{k-1}^{(j)}]$

Step 3: Move counter j to $j + 1$ and return to step 2

Once we arrive at $j = i$ we will have the values $(\theta_1^{(i)}\theta_2^{(i)}, ..., \theta_k^{(i)})$. For large i, $i \to \infty$, $(\theta_1^{(i)}\theta_2^{(i)}, ..., \theta_k^{(i)}) \xrightarrow{d} (\theta_1, \theta_2, ..., \theta_k)$

A more in depth explanation of the Gibbs sampler can be seen in Smith and Roberts (1993) and Casella and George (1992).

2.4 Data augmentation

Sometimes the models can have complicated likelihood functions which could lead to difficult calculations if handled directly. It is sometimes possible to introduce extra variables knows as auxiliary variables. These variables are not observed but, if they were, would make the calculations simpler. The variables are then treated as missing data and this is what is known as data augmentation. Tanner and Wong (1987) give examples of this in practice.

2.5 Convergence

A big difficulty that can arise using the Gibbs sampler and other MCMC methods is the ability to assess the convergence. We have seen the outcome of the Gibbs sampler is $(\theta_1^{(i)}\theta_2^{(i)}, ..., \theta_k^{(i)})$. As $i \to \infty$, the distribution of this will tend to our posterior distribution of unknowns. In reality, it is not necessary to go to ∞ as the required posterior distribution will be reached after a sufficiently large amount of iterations. The time taken to converge is known as the burn-in period.

In certain cases, convergence does not always occur. MCMCs have consecutive samples that are not independent, so we may find that the chain may spend some time in one region of the posterior. It is therefore important to analyse the MCMC sufficiently to ensure convergence has occurred and to find the size of the burn-in period. Once this is known, the samples taken in the burn-in period can be discarded and we can analyse our simulated posterior distributions.

To check convergence we can use multiple chains with different initial values and observe the trace plots. A chain that has converged will produce a trace plot with all chains moving within the same region of the parameter space with similar variation. Zellner and Min (1995) explain the properties and show practical uses of Gibbs Sampler convergence criteria.

2.6 The `rjags` package

In this project we use `rjags`, a package in R (R Development Core Team, 2004), that provides an interface from R to the JAGS library for Bayesian data analysis. JAGS, Just Another Gibbs Sampler, is a program that allows us to analyse Bayesian models using MCMC. JAGS was designed with three key features in mind: to have an engine for the BUGS language; to allow users to write their own functions, distributions and samples; and to allow experimentation with ideas in Bayesian modeling. Advice on how to use BUGS, Bayesian Inference Using Gibbs sampling, is given by Gilks et al. (1994).

Plummer (2012) provides the user manual with guidance on how to use the JAGS software and Plummer (2014) gives further details and advice on using the `rjags` package.

2.7 Regression models

2.7.1 Linear regression

Linear regression analysis allows us to test whether variables are linearly related and, if so, how strongly. The relationship can be described by the equation $Y = \alpha + \beta X$, in which Y is the variable we are predicting using the variable X. The slope parameter β represents the change in Y that is associated with one unit change in the variable X and the intercept parameter α predicts the value of Y when $X = 0$.

In cases with several predictor variables we have multiple regression. Say we have k independent variables, our $\beta_1, \beta_2, ..., \beta_k$ are our partial slope coefficients that represent the change in Y associated with one unit change in the corresponding variable $X_1, X_2, ..., X_k$. For multiple individuals, say $j = 1, ...J$, we can now describe our model by the equation

$$Y_j = \alpha_j + \beta_1 X_{1,j} + \beta_2 X_{2,j} + ... + \beta_k X_{k,j} + \epsilon_j$$

where ϵ_j is the the error associated with individual j. In a normal linear model we make the following assumptions about the error term

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

Menard (2002) describes in further detail how to examine and analyse regression models and some of their practical applications.

2.7.2 Logistic regression

In generalised linear models, we relax the first two of our assumptions about ϵ_i to allow for a broader scope of models. For a normal linear model we say that $\mu_i = \sum_{j=1}^{p} x_{i,j}\beta_j$. We introduce a quantity called the linear predictor: $\eta_i = \sum_{j=1}^{p} x_{i,j}\beta_j$. We can now introduce a variety of link functions, g, where $\eta_i = g(\mu_i)$

Considering a regression model where the error distribution is binomial, we can introduce what is called a logistic link function. Note, other link functions can be used. In this case the response variable is binary and we will obtain values $y_i = 1$ or $y_i = 0$. If we think of the mean value of y_i to be p_i then clearly a linear model $p_i = \sum \beta_j x_{i,j}$ would be inappropriate as large values of $\sum \beta_j x_{i,j}$ would lead to values greater than 1, and conversely small values of $\sum \beta_j x_{i,j}$ would lead to values smaller than 0. In this case, we need to transform our $(0, 1)$ scale to a $(-\infty, \infty)$ scale. This is achieved using logits, denoted by

$$\begin{aligned} \eta_i &= \text{logit}(p_i) \\ \eta_i &= \log\left(\frac{p_i}{1-p_i}\right) \end{aligned}$$

Here, if $p_i \to 1$ then $\eta_i \to \infty$ and if $p_i \to 0$ then $\eta_i \to -\infty$.

The inverse transformation is

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

2.7.3 Ordinal logistic regression

Sometimes in regression, we are able to measure the response variable on an ordinal scale. There are several different approaches to using ordinal models; cumulative, continuation and local. We will consider the cumulative approach, in which we compare the probability of being at or below a certain point to the probability of being beyond that point. This is used when the dependent variable represents an underlying continuous measure. The cumulative approach corresponds to the proportional odds model explained by McCullagh (1980).

We approach this method by splitting the categories, say M, of the dependent variable into $M-1$ logit equations. The outcome will be $M-1$ binary logit equations based on the comparisons 1 vs. $2-M$, $1-2$ vs. $3-M$,...,$1-(M-1)$ vs. M. From this we obtain the cumulative probability, the probability of being less than or equal to a category, that we require. Further explanation of the cumulative approach and the other types of ordinal logistic regression models are found in Fullerton (2009).

Chapter 3

The Data

3.1 *Strigamia maritima*

The Geophilomorph order of centipedes have shown to be an exception within the Arthropod phylum. In general, most groups of arthropods have a fixed number of segments for each species. Within the Geophilomorph order there is intraspecific variation in the segment numbers. The segment number of a centipede, of this order, is independent of age and is determined during embryonic development. Within this study, we are considering the number of segments of a centipede to be the number of trunk segments. A trunk segment is also know as a leg bearing segment which, from here on, we count the number of these as the number of segments. It is important to note here that the segment numbers of the *Strigamia Maritima*, shown in Figure 3.1, are always odd.

Kettle and Arthur (2000) collected data on the *S.maritima* species of Geophilomorph centipedes due to its advantageous properties. It is found in high local densities, meaning that large samples are available to provide sufficient data for analysis. Also *S.maritima* have clear sexual dimorphism in characters other than segment number. This species is restricted to coastal habitats, which simplifies the pattern in geographical variation enabling easier choice of sample sites. Finally, there have been previous studies by Horneland and Meidell (1986) on the presence of segment number variation in this species.

Figure 3.1: *S. maritima*

3.2 Data collection

The data used in my study is that of Kettle and Arthur (2000). The data were collected from ten sites along the east coast of Great Britain within a 700km span, varying from a latitude of 52.017°N to 58.642°N.These Sites can be seen in Figure 3.2. When choosing sites Kettle and Arthur ensured that there were approximately equal intersite distances and that they each had a similar degree of exposure. This avoided introducing any confounding variables. Another three further samples were taken close to the the samples already collected at Holy Island and Whitburn, which allows us to investigate any small-scale local variation that may have a significant effect.

Each sample consisted of 39 to 254 centipedes which were taken to a laboratory to determine the sex and count the number of segments. Age was not a factor that needed to be considered due to the development on the *S.maritima* being 'epimorphic'. The data collected are recorded in Table 3.1 and the corresponding site locations can be seen in Table 3.2.

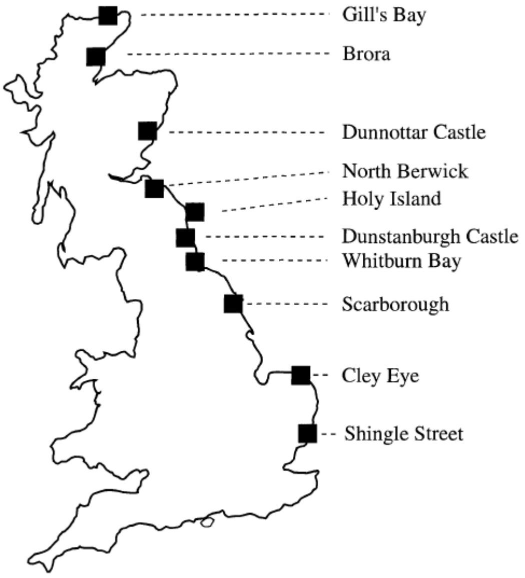

Figure 3.2: The locations of the ten main sampling sites

	Latitude	Male segment number					Female segment number					
Site	(°N)	45	47	49	Mean	Total	47	49	51	53	Mean	Total
1	58.642	14	78	0	46.70	92	13	89	7	0	48.89	109
2	57.983	10	62	0	46.72	72	7	61	4	0	48.92	72
3	56.933	2	43	14	47.41	59	0	35	18	0	49.68	53
4	56.05	0	142	50	47.52	192	0	39	23	0	49.74	62
5a	55.685	6	50	7	47.03	63	1	35	6	0	49.24	42
5b	55.673	2	62	10	47.22	74	1	28	3	0	49.13	32
5c	55.67	6	57	12	47.16	75	2	68	9	0	49.18	79
6	55.49	0	21	6	47.44	27	1	27	17	0	49.71	45
7a	54.947	0	20	10	47.67	30	0	29	20	1	49.88	50
7b	54.94	1	28	17	47.70	46	0	19	24	0	50.12	43
8	54.3	0	4	13	48.53	17	0	4	23	2	50.86	29
9	52.95	0	6	13	48.37	19	0	5	17	0	50.55	22
10	52.017	0	5	10	48.33	15	0	6	18	0	50.50	24

Table 3.1: Segment numbers collected from male and female centipedes

3.3 The data

Before carrying out any Bayesian analysis on the data, we can make some observations on the data, seen in figure 3.3. The males and females are plotted separately in order to observe the sex effects on the segment number and we have fitted a line through the data points. It is clear to see that at each latitude, the mean segment number of the female centipedes is larger than that of the males. The same pattern is evident in both sexes with the mean segment number decreasing as the latitude increases.

Site	Location
1	Gill's Bay
2	Broch (near Brora)
3	Dunnottar Castle
4	North Berwick
5a	Holy Island - Emanuel Lead
5b	Holy Island - The Basin
5c	Holy Island - Scar Jockey
6	Dunstable Castle
7a	Whitburn Bay Site 1
7b	Whitburn Bay Site 2
8	Scarborough
9	Cley Eye
10	Shingle Street

Table 3.2: The locations of the sites

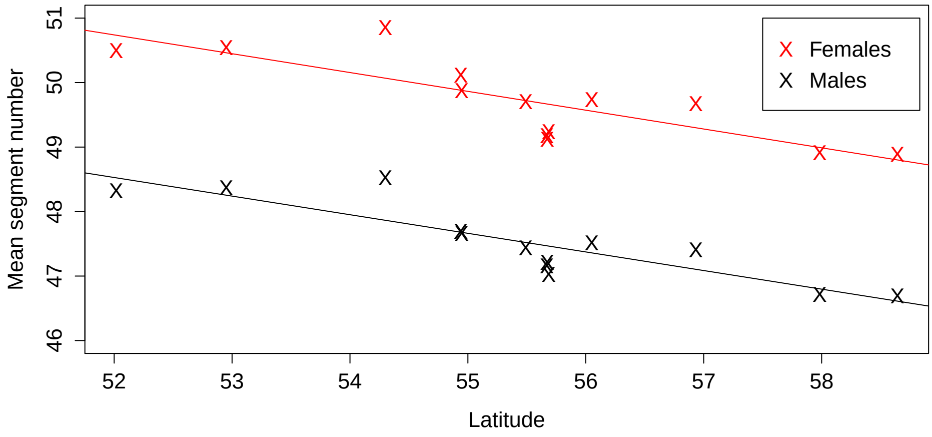

Figure 3.3: Mean segment numbers of males and females at each latitude

At this point, it is important to notice that there are a few samples collected that appear to be slightly further away than would be desired from the line of the data. This is the cluster of three points which were all collected on Holy Island. These do not follow the pattern we wish to see, suggesting there may be some other geographical factor causing the centipedes at this location to evolve differently to those on the mainland. Another outlying point to notice is the one occurring from data collected at Scarborough. Again, there may be a factor that we are not aware of interfering with the results.

Additional samples of the *S. maritima* were taken later on and observed under laboratory controlled environments to demonstrate the effects of other variables such as the temperature the centipedes are exposed to during embryonic development (Vedel et al., 2010) and a heritable component (Vedel et al., 2009)

Chapter 4

Data Analysis

4.1 Method of analysis

When analysing the data, we must consider the three different aspects involved. First of all, we must define a model to fit the data. Secondly, we must consider the inferential paradigm. We are considering Bayesian analysis of the data so we must elicit prior information for our model parameters. Finally, we need to use numerical and computational methods to calculate the posterior distributions of interest using `rjags`.

4.1.1 The model

The model that we fit to the data has several aspects. Firstly, we define our response variable $y_{i,s}$, for a centipede at site i with sex s, to have a multinomial distribution with 5 categories. The distribution has a total number of $n_{i,s}$ trials with success probabilities $p_{i,s}$. In total, our data has 26 combinations of sites and sexes, i.e 2 sexes for each of the 13 sites. We represent this by the following

$$y_{i,s} \sim \mathrm{M}_5\left(n_{i,s}, p_{i,s}\right)$$

We use $n_{i,s}$ to be the total number of centipedes observed at site i with sex s and $p_{i,s}$ is a column vector of five probabilities corresponding to a centipede presenting a particular number of segments given its characteristics. The 5 categories that our multinomial distribution has correspond to the centipedes having a certain number of segments present, from 45 through to 53 in steps of 2. It is important to note that category 1 is actually the event of having 45 or fewer segments and category 5 is actually the event of having 53 or more segments. No segment number outside the range $45 \leq N \leq 53$ was observed in the samples. N denotes the the number of segments.

In order to model these probabilities, $p_{i,s,1:5}$ for each category, we introduce the cumulative probabilities $q_{i,s,1:5}$. These represent the probability of a centipede being in each of our 5 categories. For example $q_{i,s,1} = \Pr(N \leq 45)$ and so clearly $q_{i,s,5} = \Pr(N < \infty) = 1$. We can find the values of $p_{i,s,1:5}$ by finding the difference between the consecutive cumulative probabilities, $q_{i,s,1:5}$. For example, we find $p_{i,s,3}$ by calculating $q_{i,s,3} - q_{i,s,2}$.

We fit these probabilities to a linear model using a logit link function which, as discussed in section 2.7.2, transforms our probabilities from a $(0,1)$ scale to a $(-\infty,\infty)$ scale. Our model is

$$\log\left(\frac{q_{i,s,h}}{1-q_{i,s,h}}\right) = \gamma_h + \alpha_s + \beta\left(x_i - 55\right) \tag{4.1}$$

where γ_h is our intercept parameter, α_s is our partial slope coefficient for the effects of the sex of the centipede, β is our partial slope coefficient for the effect of the site and x_i is the latitude of site i. In this model we have centered the latitude data around 55, which is roughly the center of the range of latitudes. As our γ parameter is ordered we want them to be subject to $\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4$.

4.1.2 Prior specification

In order to implement this model, it is necessary for us to specify the prior beliefs. O'Hagan et al. (2006) and Congdon (2005) discuss the complex processes and decisions involved with eliciting the prior information.

We first look at the mean of the γ's. We choose γ_3 to be the probability of having less that or equal to 49 segments. We expect half of our sample to fall either side of this parameter so we centre our beliefs around this by giving it a mean of 0. About this, we give γ_1,γ_2 and γ_4 means $-10, -5$ and 5 respectively. We can justify these allocations by considering the following model. If we had

$$\alpha_s + \beta(x_i - 55) = 0$$

then we would have

$$\text{logit}(q) = \gamma$$

Or inversely, we could say that

$$q = \frac{\exp(\gamma)}{1+\exp(\gamma)}$$

Substituting in our mean values for γ, we obtain the values q=0.00004539, 0.0066928, 0.5 and 0.993071 for means of -10, -5, 0 and 5 respectively.

As we do not want to be too informative with the prior assumptions, we give the γ's a fairly large variance of 20. This corresponds to a standard deviation of approximately 4.5 and therefore 2s.d$\approx$ 9. As our 95% intervals for our mean values are $\pm$2s.d's we obtain the 95% interval for the mean -10 to be

$$\left(\frac{\exp(-19)}{1+\exp(-19)}, \frac{\exp(-1)}{1+\exp(-1)}\right) = (5.602796e-09, 0.2689414)$$

Similarly, we obtain the intervals (8.31528e-07, 0.9820138), (0.0001233946, 0.9998766) and (0.01798621, 0.9999992) for the means -5, 0 and 5 respectively. These intervals are clearly large, as desired.

Looking now at the effects of sex, α, on the number of segments, we must first use the constraint that $\alpha_1 + \alpha_2 = 0$ in order to avoid over-parameterising the model. Again we give this a Normal distribution. We do not want to assume that the sex of a centipede is going to affect how many segments, in any direction, that the centipede will have. This is done by giving it a mean of 0.

We then give this a relatively large variance of 10 as we do not want to be too informative in the prior information. This variance of 10 corresponds to a standard deviation roughly equal to 3, giving us $2s.d \approx 6$. It follows then that we expect the values of α to fall within 6 units either side of the mean, 0. This time by considering the model when the γ and β effects are zero we get the model

$$\alpha = \text{logit}(q)$$

Again by substituting $\alpha = 6$ and $\alpha = -6$ into the inverse equation, we can find our 95% interval for q to be $q \in (0.00247, 0.9975)$ which is large, as desired.

Similarly with the effects of latitude, β, we do not want to assume anything or be too informative, so again we give this a Normal distribution with a mean of zero and a fairly large variance of 25. This corresponds to 2s.d's=10, however we also have the element of centering occurring for this parameter, i.e. we multiply it by $(x_i - 55)$. Our latitudes fall roughly within $\pm 3^o$ of 55^o so this combines with our standard deviation to give us an interval of ± 30. This is very large and obtains the interval $q \in$(9.357623e-14, 1).

In summary, our prior distributions are

$$\begin{aligned}
\gamma_1 &\sim \text{N}(-10, 20)\\
\gamma_2 &\sim \text{N}(-5, 20)\\
\gamma_3 &\sim \text{N}(0, 20)\\
\gamma_4 &\sim \text{N}(5, 20)\\
\beta &\sim \text{N}(0, 10)\\
\alpha_1 &\sim \text{N}(0, 25)
\end{aligned}$$

4.1.3 The `rjags` specification

Now that we have specified the model and the prior information that we wish to use we can compute our posterior distributions and summaries using `rjags`. We do this using the following code to create our model file

```
model

{
        for (i in 1:26) {
                y[i,]~dmulti(p[site[i],sex[i],],n[i])
                         }

        for (st in 1:13) {
            for (sx in 1:2) {
                 p[st,sx,1]<-q[st,sx,1]
                 p[st,sx,2]<-q[st,sx,2]-q[st,sx,1]
                 p[st,sx,3]<-q[st,sx,3]-q[st,sx,2]
                 p[st,sx,4]<-q[st,sx,4]-q[st,sx,3]
                 p[st,sx,5]<-1-q[st,sx,4]
                        for (h in 1:4) {
                                logit(q[st,sx,h])<-gamma[h]+alpha[sx] +beta*(x[st]-55)
                                       }
                              }
                            }
        gamma0[1]~dnorm(-10,0.05)
        gamma0[2]~dnorm(-5,0.05)
        gamma0[3]~dnorm(0,0.05)
        gamma0[4]~dnorm(5,0.05)
        gamma[1:4]<-sort(gamma0)
        alpha[1]~dnorm(0,0.1)
        alpha[2]<-0-alpha[1]
        beta~dnorm(0,0.04)
}
```

We then read this model file into R using the `jags.model` function. In order to define the burn-in period we use the command `update` . To create our MCMC sample that summarises the posterior distribution we use the `coda.samples` function. This also provides convergence diagnostics that allow us to check that our output is valid to analyse.

4.2 Assessing the convergence of the sampler

Having fitted a model to my data using MCMC, it is important to ensure good mixing and convergence of the chains has occurred before continuing. This has already been done using the `coda.samples` function in R but we must also look at the trace plots. An example of this can be seen in Figure 4.1.We have used two parallel chains, using colours black and red for clarity, to observe convergence from different starting values. Clearly, for each γ, the chains have not moved much from the initial value, showing an appropriate burn in period has been used. It is also evident that both chains in each graph are moving within the same region of the parameter space with similar variation, suggesting the mixing is good.

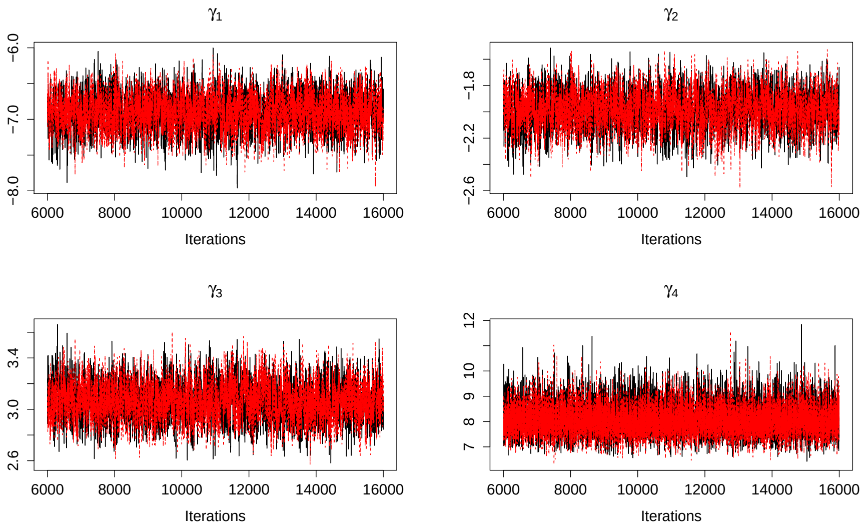

Figure 4.1: Trace plots of the posterior information of the γ's

This is apparent with trace plots of the other parameters, suggesting the simulated posterior distributions all converge and are sufficient to continue to analyse.

4.3 Assessing the posterior information

Using the `rjags` simulated values, we can plot posterior densities which can then be compared to those of the priors. This will allow us to see how the prior beliefs have been updated by the data.

Figure 4.2 shows the density plots of $\gamma_{1:4}$. The posterior densities for each of the parameters clearly have a much smaller variance than the prior densities. Clearly the posterior information is more precise than the prior and we are more certain about the values of the parameters. For each graph we can see a slight increase in our belief about the mean value from -10,-5,0,5 to -6.902707, -1.984655, 3.065809 and 8.112955 respectively.

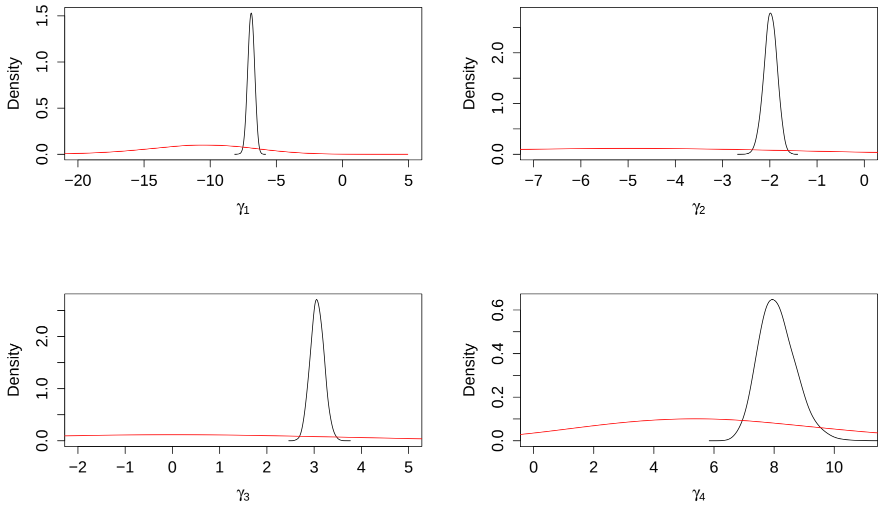

Figure 4.2: Prior (red) and posterior (black) density plots of the γ's

Figure 4.3 shows the density of α_1, which corresponds to the sex effect, specifically the male sex effect. Again, there is a significant reduction in the variance from the prior to the posterior. Our large prior variance of 10, i.e 2s.d $\approx \pm 6$, has therefore not influenced our results and we are now more certain about our beliefs.

We can see that the mean has increased from 0 to 2.729, showing that the sex of the centipede does in fact affect the outcome of the number of leg bearing segments that a centipede will develop. Due to the structure of the model, having a positive mean translates as the centipede being more likely to have a smaller number of segments. As α_1 is the male sex effect we can say that a male centipede is more likely to have fewer leg bearing segments than a female centipede at the same latitude. As we set the constraint $\alpha_1 = -\alpha_2$, the female sex effect will have a negative mean and will therefore mean the female centipedes are more likely to have a larger number of leg bearing segments than male centipedes at the same latitude.

Figure 4.4 shows the density of β, the effect of latitude on the number of segments present in the centipede. Again the variation has decreased from prior to posterior so we are now more certain about the value of β. Assigning the variance of 25 has turned out to be very large and has meant that our prior distribution has not had too much of an effect on the posterior distribution and the information we use is coming primarily from the data.

The mean of β has slightly increased from 0 to about 0.707. Similarly to before, having a positive mean for our parameter shows that as the latitude increases the centipede becomes more likely to have fewer leg bearing segments than a centipede found at a lower latitude.

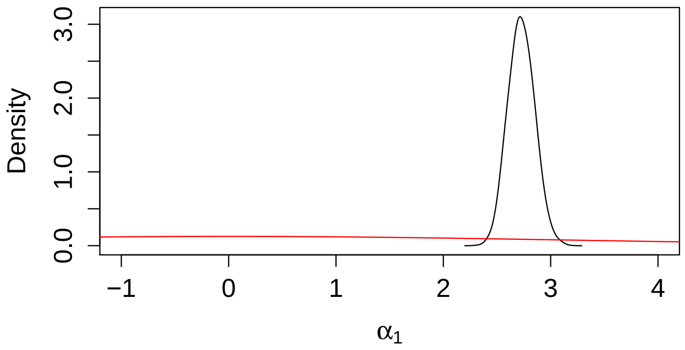

Figure 4.3: Prior (red) and posterior (black) density plot of α_1, the sex effect

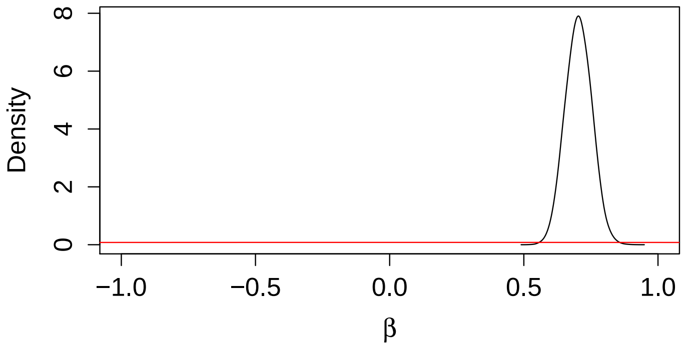

Figure 4.4: Prior (red) and posterior (black) density plot of β, the latitude effect

4.4 Bivariate posterior density plots

Using R, we produced bivariate posterior density plots in order to observe the dependence between the parameters. The plots for the neighboring γ parameters are shown in Figure 4.5. As we have simulated lots of data these plots are very clustered and not necessarily clear so, using a function in R shown in the listing below, we estimated the bivariate density in order to produce the smooth contour plots also shown in Figure 4.5. We originally expected to see the γ variables all having a positive correlation, with the strongest correlation occurring with neighboring variables. However, this has not been apparent in the resulting graphs.

Listing 4.1: Bivariate kernel smoother breaklines

```
bivardens<-function(xlims,ylims,data,adj=1)
{
  n<-length(data[,1])
  W<-solve(var(data))*adj
  xvals<-seq(xlims[1],xlims[2],length.out=101)
  yvals<-seq(ylims[1],ylims[2],length.out=101)
  dens<-matrix(0,nrow=101,ncol=101)
  for (ix in 1:101) {
    write(ix,file="")
    x<-data[,1]-xvals[ix]
    for (iy in 1:101) {
```

```
   y<-data[,2]-yvals[iy]
   for (i in 1:n) {
    xy<-c(x[i],y[i])
    dim(xy)<-c(2,1)
    dens[ix,iy]<-dens[ix,iy]+exp(-(t(xy)%*%W%*%xy))
    }
   }
  }
 stepx<-xvals[2]-xvals[1]
 stepy<-yvals[2]-yvals[1]
 dens<-dens/(sum(dens)*stepy*stepx)
 result<-list(x=xvals,y=yvals,density=dens)
 return(result)
}
```

From Figure 4.5, we observe that the plot for γ_1 vs .γ_2 fits to what we expect. The plots for γ_2 vs. γ_3 and γ_3 vs. γ_4 do not behave as we would wish. At this point we believe that the problem lies with γ_3 and γ_4. These correspond to the probability of having less than or equal to 49 segments and less than or equal to 51 segments. Looking back to the data collected, we can see that we only have samples of females with more than 49 segments. One concern that arose previously in the study was the abnormal figures collected at Holy Island. In particular at this site, it can be seen that we have low numbers of female centipedes taken at this point relative to the numbers collected at nearby sites. Perhaps there is an unknown geographical factor that is causing abnormal results.

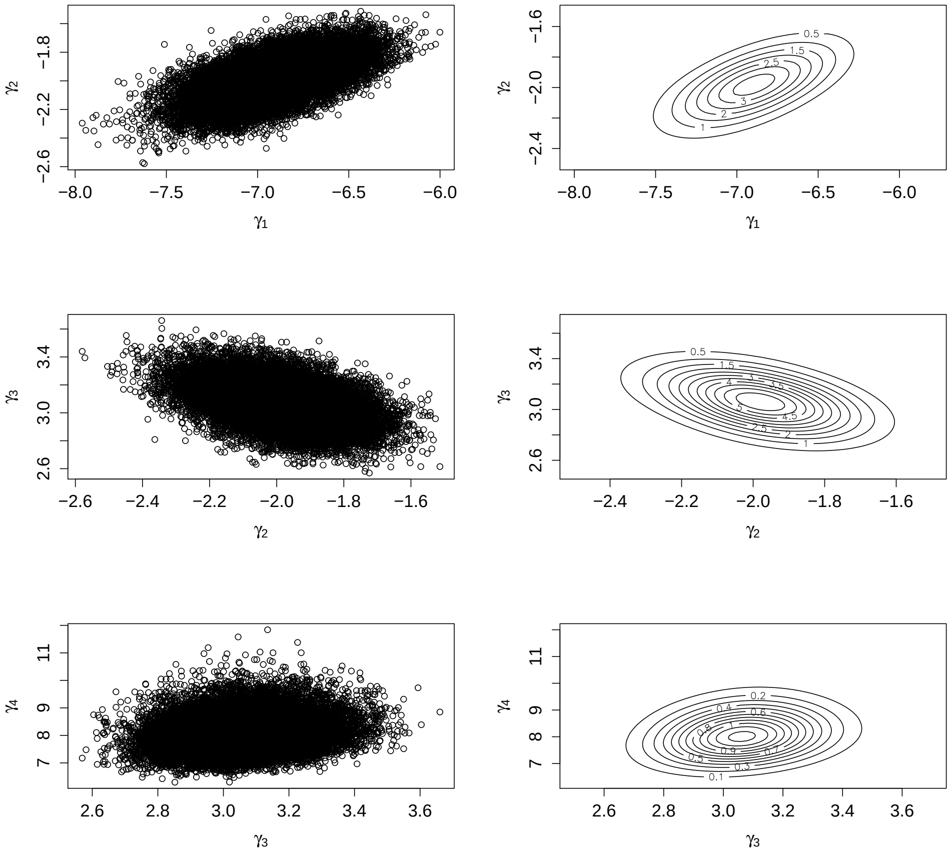

Figure 4.5: Bivariate plots of the γ parameters

Chapter 5

Improving the Model

5.1 Adding a random site effect

Having come across an unexpected result, it seems appropriate to adapt the model to include a random effect that accommodates this deviation from the simple model. Our model becomes

$$log\left(\frac{q_{i,s,h}}{1-q_{i,s,h}}\right) = \gamma_h + \alpha_s + \beta\left(x_i - 55\right) + r_i \tag{5.1}$$

$$r_i \sim \mathrm{N}(0, \tau_{site}^{-1})$$

$$\tau_{site} \sim \mathrm{Ga}(1.1, 0.3)$$

where r_i is a random site effect for site i. We assign the same prior information to the parameters in equation (5.1) as before for equation (4.1) and we have given the random effect a Normal distribution with a zero mean. Any nonzero mean will be absorbed by the fixed effects in the model. For the previous parameters, we assumed constant variation at each latitude. For this random effect we have given the precision a $\mathrm{Ga}(1.1, 0.3)$ distribution to relax the straight-line assumption of our previous model.

5.2 Assessing the model

By adapting our previous R code, we can now produce plots that allow us to analyse our new model. The γ trace plots are shown in Figure 5.1. Again we use two chains in our simulation in order to assess the convergence and mixing of the sample. From the figure we see the two separate chains, in red and black, are both mixing well within the same region of the parameter space. This is apparent in all of the posterior trace plots, suggesting to us that our burn in period and thinning are sufficient and we can continue to analyse the data.

5.3 Assessing the posterior information

Figure 5.2 shows the posterior distribution of both this model and our previous model, for comparison. We can see that our new posterior, blue, has a smaller variance than the prior distribution, but not smaller than the variance of our previous posterior distribution, black. This is due to the fact that the random effect in our new model creates deviation

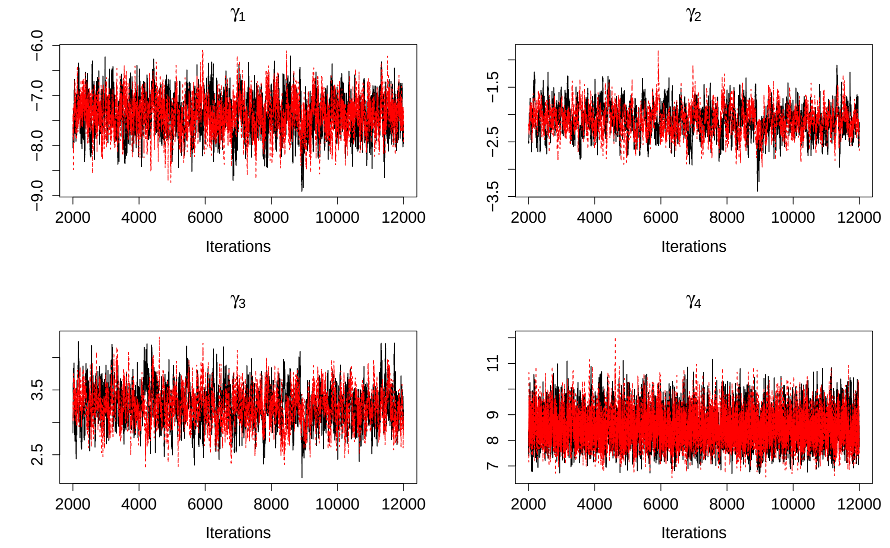

Figure 5.1: Trace plots of $\gamma_{1:4}$

from the straight-line assumption. This has the effect of relaxing the structure of the model but making our beliefs more realistic.

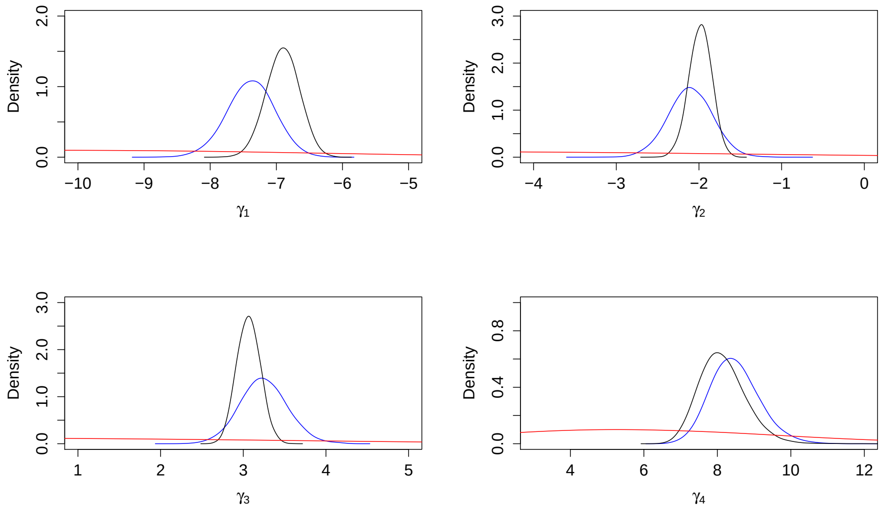

Figure 5.2: Prior (red), posterior (Blue) and posterior from previous model (black) density plots for the intercept parameters, $\gamma_{1:4}$

Figure 5.3 shows that, for the sex effect, our posterior distribution, blue, is quite similar to that of the previous,black. This we expected to see as the random effect that has been added to the model is only being affected by site and not on sex. So as before, being male means the centipede is more likely to have fewer segments.

Unlike our α parameter, the random effect has had a big impact on the distribution of our β parameter. This was to be expected as the beta parameter is related to the site of the sample. Figure 5.4 shows us that our posterior distribution, blue, has a smaller variance than that of the prior but again not as small as that of the previous posterior, black. However, our new posterior distribution is perhaps a better representation of the

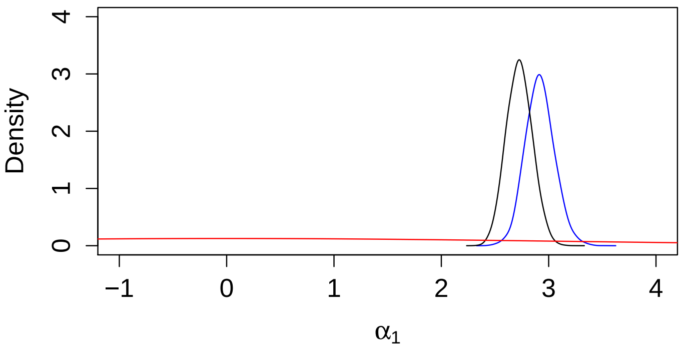

Figure 5.3: Prior (red), posterior (blue) and posterior from previous model (black) density plots of α_1, the sex effect

data. The mean has stayed almost the same as the one obtained from our previous model so we still expect to find centipedes with fewer segments as we observe a sample from a site of higher latitude.

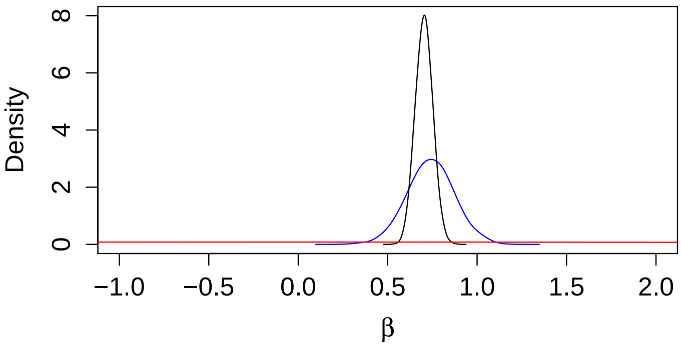

Figure 5.4: Prior (red), posterior (blue) and posterior from previous model (black) density plots of β, the latitude effect

5.4 The variance component

In our new model, we introduced a random site effect which had a precision τ_{site}. Using this precision we can observe the deviation of our data by noting that τ_{site}^{-1} is our variance component. In order to observe the effects of this component we wanted a density plot of the prior and posterior distributions. It is important to note that the posterior data is simulated and so plotting its density with a smoothing kernel will produce negative values. This is undesirable as τ_{site} has a gamma distribution which has the property that it returns positive real numbers.

To overcome this, we must first transform the data using logarithms and then use a kernel smoother to obtain the density of the transformed data. In order to get our data back into the desired form we must exponentiate the x values obtained from the kernel smoother. In order to obtain our new distribution we must note the following about how we transformed x.

$$\tau_{site} = e^x$$

and so

$$\frac{d\tau_{site}}{dx} = e^x = \tau_{site}$$

Now, using

$$f_\tau(\tau_{site})d\tau_{site} = f_x(x)dx$$

we can see that the following equations gives us our density

$$f_\tau(\tau_{site}) = {}^{f_x(x)}\!/\!_{\frac{d\tau_{site}}{dx}} = {}^{f_x(x)}\!/\!_{\tau_{site}}$$

Having transformed our data we can produce the plot shown in figure 5.5. Clearly, the posterior variance has decreased so we are now more certain about the value of τ_{site}. Looking at our R output we can see that it has actually decreased from 12.222 to 0.7613736 and that our mean value has decreased from 3.6667 to 1.893032. Having observed this parameter we can say that any deviation of our data from the line of regression is not just due to the random sampling but also due to the individual sites at each latitude.

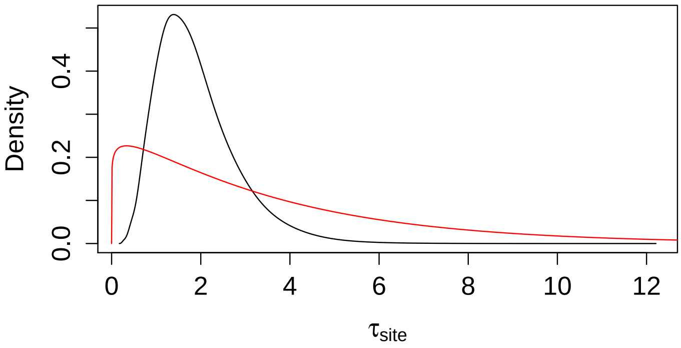

Figure 5.5: Prior (red) and posterior (black) density plot of τ_{site}

5.5 Bivariate density plots

Using the R code from Section 4.4, we again produced bivariate plots and contour plots. These are shown in Figure 5.6. We now see the desired effect we were previously looking for. It is clear to see that the neighboring γ parameters are dependent on one another. This suggests to us that adding the random effect to our model has resolved the problem we experienced previously and leads us to believe the model has been improved.

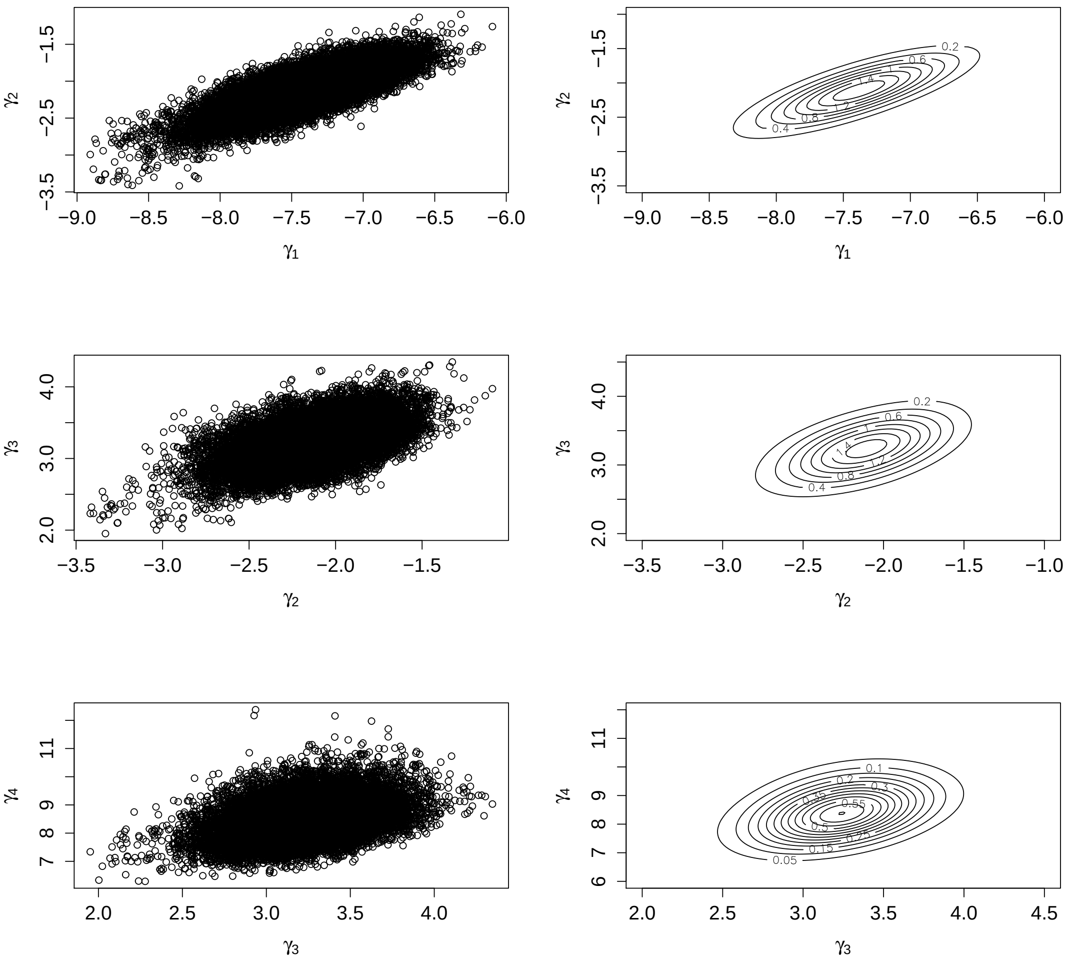

Figure 5.6: Bivariate posterior scatter and contour plots of the γ variables with the random variable

Chapter 6

Addition of a Further Random Effect

6.1 The cluster effect

Having seen that we do indeed have variation between the different sites, it seems sensible to investigate the possibility of sites within a cluster, where more than one site appears within the same region, having variation. This involves us adding a further random effect to our model to get

$$\log\left(\frac{q_{i,s,h}}{1-q_{i,s,h}}\right) = \gamma_h + \alpha_s + \beta\left(x_i - 55\right) + r_i + w_{\text{cluster}}$$

where

$$w_{\text{cluster}} \sim \mathrm{N}(0, \tau_{\text{cluster}}^{-1})$$

Before we assign prior information to the random cluster effect, w_{cluster}, we must consider the logistics of variance components.

6.2 Learning about the model

Adding random variables to the model produces different outcomes depending on how many variables we add. Supposing that we only have one random effect, the variation of the deviation from the line, say ϵ, for the model is given to this random variable. If we added another random effect so we now have two random variance components, our ϵ will be split between the two.

For the model with one random effect, this deviation can be represented as

$$\mathrm{var}(\epsilon) = \frac{1}{\tau_{\text{site}}}$$

Introducing a model that has two variance parameters, we need to note that we now have

$$\begin{array}{rcl} \mathrm{var}(\epsilon) & = & {}^{1}\!/\tau_{\text{site}} + {}^{1}\!/\tau_{\text{cluster}} \\ \Longrightarrow\ \tau_{\text{total}} & = & {}^{1}\!/[\mathrm{var}(\epsilon)] \end{array}$$

We now have to adapt our model to accommodate for our variance of deviation being assigned to two random effects. In our model specification, we have introduced a variable

τ_{total} which behaves as though it is the only variance component in our model. So we assign this component the same Ga$(1.1, 0.3)$ distribution that we assigned to our model with one random effect in Section 5. The plot in Figure 6.1 shows the prior and posterior density plots for τ_{total}.

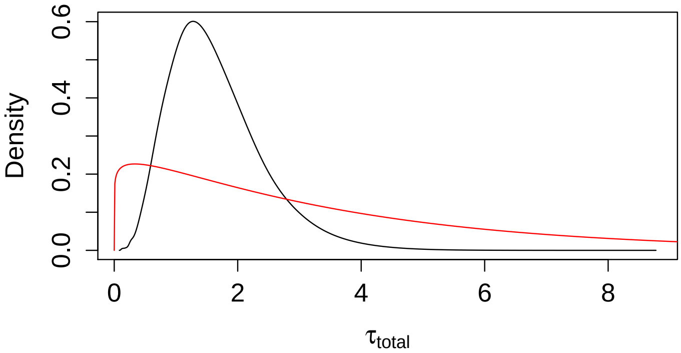

Figure 6.1: Prior (red) and posterior (black) density plot of τ_{total}

From Figure 6.1 we can see that the variance of the posterior distribution has decreased from that of the prior distribution, so we are more certain about our value. More specifically, if we compare this plot to Figure 5.5 in Section 5.4 we can see that our distributions are in fact the same, due to the total variance of deviation from the line remaining the same.

6.3 Prior specification

In order to observe the effects of our individual random effect parameters we must define the ratio of the total variance that the two variance components take. To do this, we have introduced another parameter δ that has a Beta(1.5,1.5) distribution.

Using the distributions of δ and τ_{total} we can define our variance components to be

$$\begin{aligned} \tau_{\text{site}} &= \frac{\tau_{\text{total}}}{\delta}, \\ \text{and} \quad \tau_{\text{cluster}} &= \frac{\tau_{\text{total}}}{1-\delta} \end{aligned}$$

Using these ratios, we can evaluate our joint prior distributon for τ_{site} and τ_{cluster}. We first rearrange to find $\tau_{\text{total}}(\tau_t)$ and δ in terms of $\tau_{\text{site}}(\tau_s)$ and $\tau_{\text{cluster}}(\tau_c)$. We obtain

$$\begin{aligned} \tau_t &= \frac{\tau_s \tau_c}{\tau_s + \tau_c}, \\ \text{and} \quad \delta &= \frac{\tau_c}{\tau_s + \tau_c} \end{aligned}$$

Using the prior distrions $\tau_t \sim \text{Gamma}(a, b)$ and $\delta \sim \text{Beta}(c, d)$ we find the joint prior density to be

$$\frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} \delta^{c-1}(1-\delta)^{d-1} \frac{b^a}{\Gamma(a)} \tau_t^{a-1} e^{-b\tau_t}$$

Substituting in τ_t and δ in terms of τ_s and τ_c, the joint prior distribution of τ_s and τ_c is

$$\frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)}\left(\frac{\tau_c}{\tau_s+\tau_c}\right)^{c-1}\left(\frac{\tau_s}{\tau_s+\tau_c}\right)^{d-1}\frac{b^a}{\Gamma(a)}\left(\frac{\tau_s\tau_c}{\tau_s+\tau_c}\right)^{a-1} e^{-b\frac{\tau_s\tau_c}{\tau_s+\tau_c}}|\mathrm{J}|$$

where $|\mathrm{J}|$ is the determinant of the Jacobian matrix which is

$$|\mathrm{J}| = \begin{vmatrix} \frac{\tau_s}{(\tau_s+\tau_c)^2} & \frac{-\tau_c}{(\tau_s+\tau_c)^2} \\ \frac{\tau_s^2}{(\tau_s+\tau_c)^2} & \frac{\tau_c^2}{(\tau_s+\tau_c)^2} \end{vmatrix} = \frac{\tau_s\tau_c}{(\tau_s+\tau_c)^3}$$

The joint prior density of τ_s and τ_c is therefore

$$\frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)}\frac{b^a}{\Gamma(a)}\frac{\tau_s^{a+d-1}\tau_c^{a+c-1}}{(\tau_s+\tau_c)^{a+c+d}}\exp\left\{\frac{-b\tau_s\tau_c}{\tau_s+\tau_c}\right\}$$

After checking that all of the convergence criteria were met we produced the graphs in Figure 6.2 and Figure 6.3 which show the prior and posterior distributions of our variance components.

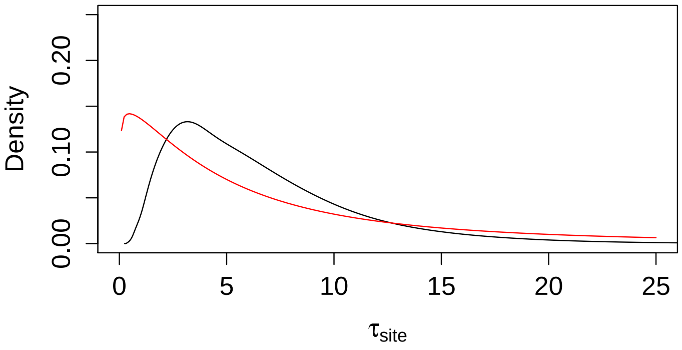

Figure 6.2: Prior (red) and posterior (black) density plot of τ_{site}

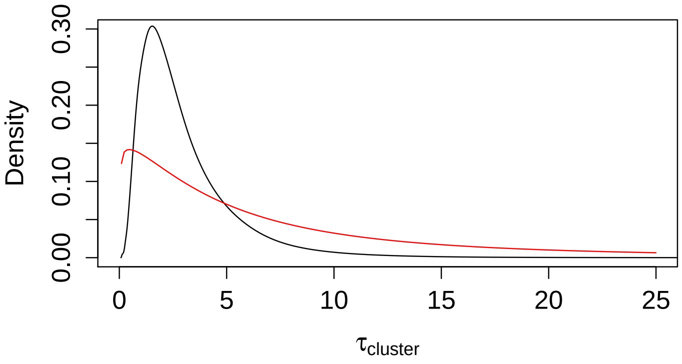

Figure 6.3: Prior (red) and posterior (black) density plot of τ_{cluster}

From Figure 6.3 we can see that there has been a big reduction in our variance from the prior to the posterior distribution, so we have learnt a bit about this parameter. Our mean value has changed from 12.55 to 4.253238. This supports our theory that sites within a cluster do have similar traits and so we do not wish to count them as individual sites.

We can see from Figure 6.2 the variance of τ_{site} has not decreased as much as τ_{cluster} so we have learnt a little bit about this but not quite as much as τ_{cluster}. This is due to the sites in each cluster no longer counting as individual sites. The mean has decreased from 12.55 to 6.50076.

Chapter 7

The Effects of a Heritable Component in the Development of Segment Number in Offspring

7.1 Introduction

Having drawn conclusive results from the latitudinal data, it seems appropriate to observe another factor that may influence the segment number variation of the *S.maritima*. We have chosen to look at heritability.

Vedel et al. (2009) obtained 204 clutches with their mothers from the field site at Brora, Scotland. The samples were then preserved in tubes with 70% ethanol and stored under temperature controlled conditions, at 4^oC, to ensure the preservation of the hatchlings soft cuticle. The hatchlings were stained with methanol blue to distinguish better the segment and genital structures and pictures were obtained. From here, it was possible to analyse the hatchlings and determine the gender and number of segments.

Size of mother	Male segment number						Female segment number					
	43	45	47	49	Mean	Total	47	49	51	53	Mean	Total
47	1	37	45	0	46.06	83	49	34	2	0	47.90	85
49	0	53	260	32	46.88	345	42	259	16	1	48.85	318
51	0	1	15	7	47.52	23	1	9	5	0	49.53	15

Table 7.1: Size of the offspring compared to the size of the parent

The raw data obtain can be seen in Table 7.1. We have then plotted these to visualise any initial observations that may be present. This plot can be seen in Figure 7.1. This shows that, generally, offspring present higher segment numbers when their parents have higher segment numbers.

7.2 Analysis

In order to draw any conclusions from this data, we must use the same method of analysis used in Chapter 4.

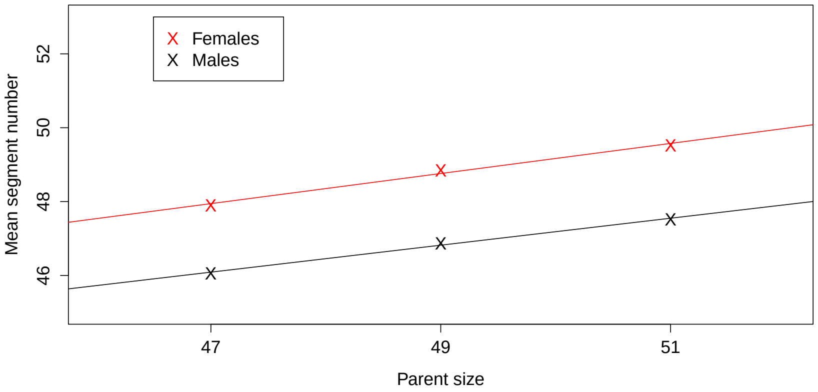

Figure 7.1: Mean segment number of offspring

7.2.1 The model

We have the response variable $y_{m,s}$ for a centipede of sex s from a mother from group m. There are three different sizes for m, $m_1 = 47$,$m_2 = 49$ and $m_3 = 51$. The response variable is given a multinomial distribution with 6 categories, a total of $n_{m,s}$ trials with success probabilities $p_{m,s}$. This data has a total of 6 combinations of size of parent with sex of offspring, i.e. 2 sexes for each of the three sizes of parent centipede. So

$$y_{m,s} \sim M_6\left(n_{m,s}, p_{m,s}\right)$$

Our $n_{m,s}$ represents the total number of offspring observed to be sex s from a mother from group m and $p_{m,s}$ to be a column vector of length six, corresponding to the probability that the offspring will present a certain number of segments. In this data set, we have offspring with segment numbers in the range $43 \leq N \leq 53$, remembering that centipedes will only present an odd number of leg bearing segments and that N =number of leg bearing segments.

Again, we have to introduce the cumulative probabilities $q_{m,s,1:6}$. Similar to before, $q_{i,s,1} = \Pr(N \leq 43)$,...,$q_{i,s,6} = \Pr(N < \infty)$. We can then work out the values of $p_{m,s,1:6}$ by subtracting consecutive cumulative probabilities. For example, $p_{m,s,5} = q_{m,s,5} - q_{m,s,4}$.

We transform the cumulative probabilities using the logit link function. Our model is

$$\log\left(\frac{q_{m,s,h}}{1-q_{m,s,h}}\right) = \gamma_h + \alpha_s + \beta\left(x_m - 49\right) \qquad (7.1)$$

where γ_h is our intercept parameter, α_s is our partial slope coefficient for the effects of the sex of the offspring, β is our partial slope coefficient for the heritable effect and x_m is the number of leg bearing segments of the parent in group m. In this model, we have centered the size of the parent data around 49, which is the center of the range of parent sizes. As our γ parameter is ordered we want them to be subject to $\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4 < \gamma_5$.

7.2.2 Prior specification

When assigning prior information to these parameters, we use similar methods for our choices as in Section 4.1.2 for model 4.1.

We assign $\gamma_{1:6}$ a Normal distribution, with means -10, -5, 0, 5, 10 and variances 100, 20, 20, 20, 100 respectively. In this case the variance of each γ is not the same. This is because we were more certain about the values in the middle of the range of segment numbers, as opposed to those at the ends. We also know that it takes a large change in the logit to produce a small change in the probability near the ends.

For β we have a normal distribution with mean 0 and variance 20, so we are being vague and not assuming anything in our prior. For similar reasons, we give α_1 a normal distribution with mean 0 and variance of 10. We also have the constraint again that $\alpha_1 + \alpha_2 = 0$ so we are not over-parameterising the model.

So in summary

$$\begin{aligned}
\gamma_1 &\sim \mathrm{N}(-10, 100) \\
\gamma_2 &\sim \mathrm{N}(-5, 20) \\
\gamma_3 &\sim \mathrm{N}(0, 20) \\
\gamma_4 &\sim \mathrm{N}(5, 20) \\
\gamma_5 &\sim \mathrm{N}(10, 100) \\
\beta &\sim \mathrm{N}(0, 20) \\
\alpha_1 &\sim \mathrm{N}(0, 10)
\end{aligned}$$

7.2.3 The `rjags` specification

We can now create our model file to compute our values using the following code

```
model

{
        for (i in 1:6) {
                y[i,]~dmulti(p[mother[i],sex[i],],n[i])
                          }
                for (mc in 1:3) {
                            for (sx in 1:2) {
                 p[mc,sx,1]<-q[mc,sx,1]
                 p[mc,sx,2]<-q[mc,sx,2]-q[mc,sx,1]
                 p[mc,sx,3]<-q[mc,sx,3]-q[mc,sx,2]
                 p[mc,sx,4]<-q[mc,sx,4]-q[mc,sx,3]
                 p[mc,sx,5]<-q[mc,sx,5]-q[mc,sx,4]
                 p[mc,sx,6]<-1-q[mc,sx,5]
                        for (h in 1:5) {
                                logit(q[mc,sx,h])<-gamma[h]+alpha[sx]+beta*(x[mc]-49)
                                        }
                                      }
                          }
        gamma0[1]~dnorm(-10,0.01)
        gamma0[2]~dnorm(-5,0.05)
        gamma0[3]~dnorm(0,0.05)
        gamma0[4]~dnorm(5,0.05)
        gamma0[5]~dnorm(10,0.01)
        gamma[1:5]<-sort(gamma0)
        alpha[1]~dnorm(0,0.1)
        alpha[2]<-0-alpha[1]
        beta~dnorm(0,0.04)
}
```

7.3 Assessing the convergence

It is important for us to again assess the convergence and fit of this model before continuing to use MCMC to analyse the data. Some of the trace plots obtained from the data are shown in figure 7.2. These clearly show that our model is mixing and converging well and that it is sufficient to analyse.

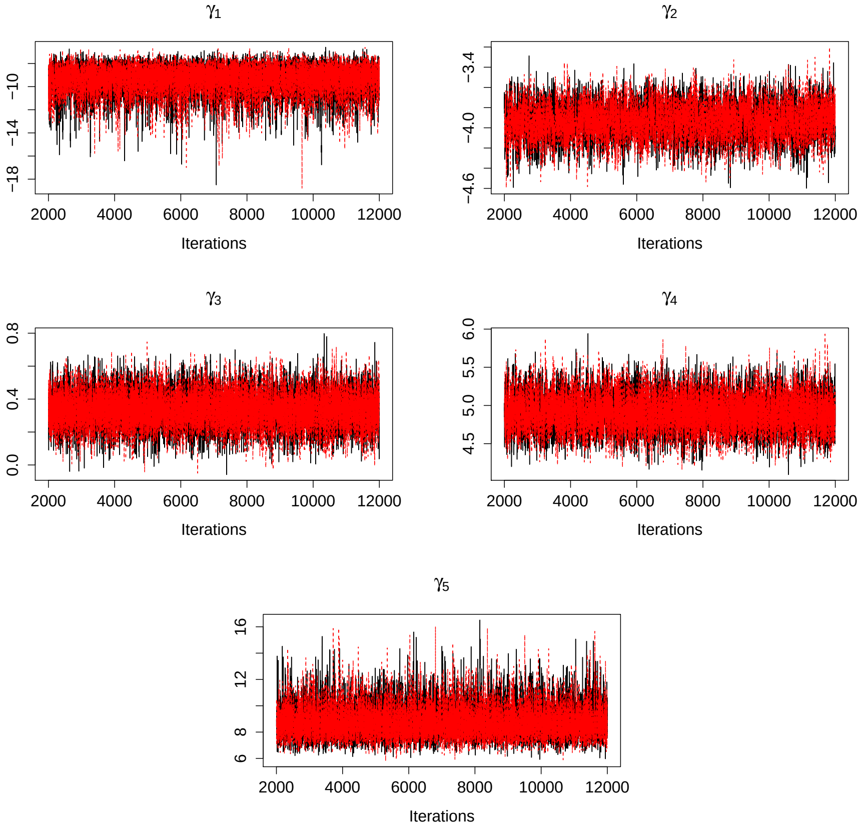

Figure 7.2: Trace plots of the γ's

7.4 Assessing the posterior distributions

From Figure 7.3, we can see that for $\gamma_{1:5}$ the variance of the posterior distribution is a lot smaller than the variance of the prior distribution. This shows that we have learnt from the data and we are more certain of our results. For each plot, the mean has stayed roughly the same. For $\gamma_{1:5}$, we find the means are -9.442857, -3.935069, 0.3365787, 4.914731 and 8.823271, respectively.

Looking at Figure 7.4, it is clear to see that the variance of the male sex effect, α_1, has reduced a lot from the prior distribution to the posterior distribution so we have learnt about α_1 from our data. Using R, we find that it has decreased from 10 to 0.01221174. We also find that the mean has increased from 0 to 2.101575, showing that a male centipede is more likely to present fewer leg bearing segments than a female centipede.

Figure 7.5 shows that the variance of the posterior distribution for β, the effect of the size of the parent centipede, has decreased from the variance of the prior distribution, from 20 to 0.006506826. Therefore, we are now more certain about the value of our

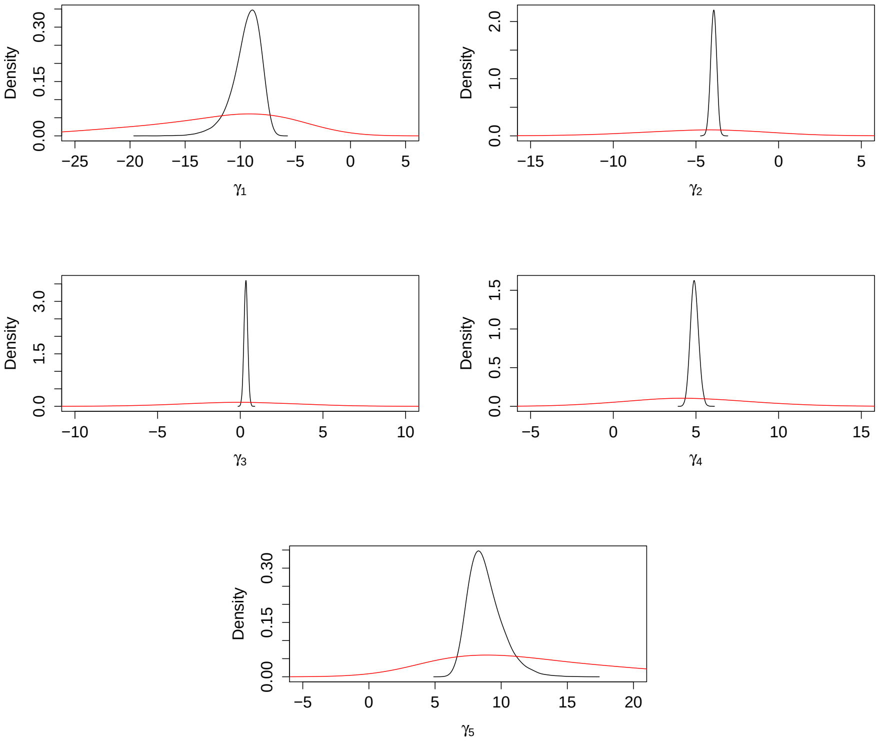

Figure 7.3: Prior (red) and posterior (black) density plots of $\gamma_{1:5}$, the intercept parameters

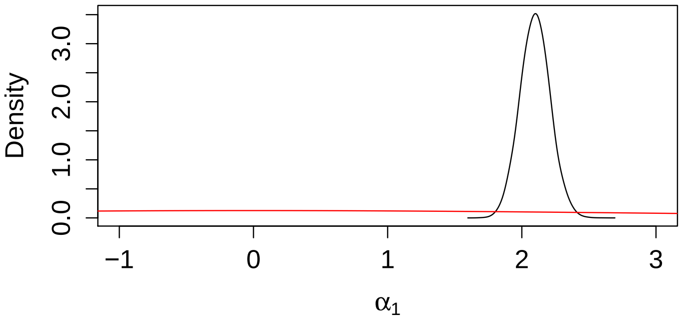

Figure 7.4: Prior (red) and posterior (black) density plot of α, the sex effect

parameter. It is also evident that the mean has decreased from 0 to -0.9168345. Due to the structure of the model, this tells us that the offspring with parents presenting higher segment numbers are more likely to have more leg bearing segments than those from a mother presenting fewer leg bearing segments.

7.5 Is this really due to a genetic effect?

There does appear to be a genetic effect but the mean increase in segment number of the offspring for an increase of 2 in the mother's segment number is less than 2. Perhaps this could be due to an environmental effect or potentially just "regression to the mean".

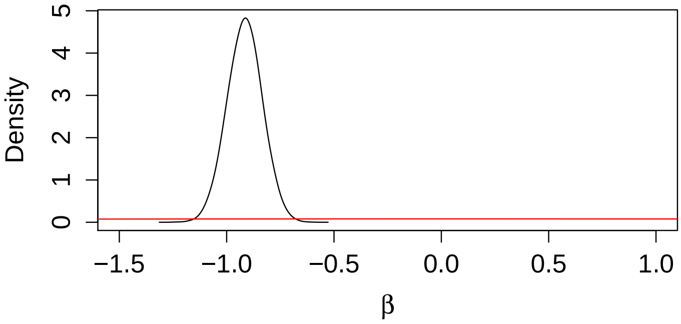

Figure 7.5: Prior (red) and posterior (black) density plot of β, the parent effect

Regression to the mean occurs when there is random variation in observed data around the true mean. This is a common occurrence when observing the effects of genetics as seen in Atkinson et al. (2010) and other studies such as Maher and Mountain (2009). Barnett et al. (2005) explain regression to the mean in further detail and give methods to overcome it.

To explore the possibility that there is an environmental effect occurring, we can compare the distribution of the data obtained in this experiment under laboratory conditions, Table 7.1, with that of Kettle and Arthur (2000), Table 3.1, for the particular observations obtained at Broch (near Brora). We assume that the mothers are a random sample of Bora females and that survival to adulthood is independent of segment number.

7.6 Re-analysing the data

The data we are now going to be analysing can be seen in Figure 7.2. We can see that the means obtained are almost identical for the same sex in each condition but the females are on average larger than the male centipedes.

Conditions	Male segment number						Female segment number					
	43	45	47	49	Mean	Total	47	49	51	53	Mean	Total
Wild	0	10	62	0	46.72	72	7	61	4	0	48.72	72
Laboratory	1	91	320	39	46.76	451	92	302	23	1	48.679	418

Table 7.2: Samples from the site at Brora

7.7 Analysis

7.7.1 The model

In this case we denote our response variable as $y_{c,s}$, where c denotes the condition the centipede develops in and s denotes its sex, male or female. There are two possible choices for the condition which are wild or laboratory which we assign 1 and 2 respectively. We give the response variable a multinomial distribution with 6 categories, corresponding to the range of outcomes which are segment numbers ranging from 43 to 53, remembering that segment numbers are only every odd. We have the success probabilities $p_{c,s}$ and a total number of $n_{c,s}$ trials. So we have

$$y_{c,s} \sim \mathrm{M}_6(n_{c,s}, p_{c,s})$$

The number of centipedes observed with a certain number of segments bred in a particular condition is represented by $n_{c,s}$. Our $p_{c,s}$ is a column vector of length 6, corresponding to the probability that the offspring will present a certain number of segments.

We use the difference between consecutive cumulative probabilities $q_{c,s,1:6}$ to find our values of $p_{c,s,1:6}$. Similar to before, $q_{c,s,1} = \Pr(N \leq 43)$,...,$q_{c,s,6} = \Pr(N < \infty)$. So, for example, we can find $p_{c,s,4} = q_{c,s,4} - q_{c,s,3}$. As before we must transform these with the logit link function and we obtain the model

$$\log\left(\frac{q_{c,s,h}}{1 - q_{c,s,h}}\right) = \gamma_h + \alpha_s + \beta_c \tag{7.2}$$

where γ_h is our intercept parameter, α_s is our partial slope parameter for the effects of the sex of the centipede and β_c is the partial slope parameter for the effects of the conditions during embryonic development. Again, in our model, we want our γ parameters to be subject to $\gamma_1 < \gamma_2 < \gamma_3, \gamma_4 < \gamma_5 < \gamma_6$.

7.7.2 Prior specification

We use similar methods for our prior specification here as we did in Section 4.1.2. We assign the same prior information for γ and α as we did in Section 7.2.2 for model 7.1.

In this model, β is the effect of the conditions during embryonic development. We have two different conditions and so set $\beta_1 + \beta_2 = 0$ so as not to over-parameterise our model. We then give β_1 a Normal distribution with a mean of 0 and a variance of 20. This ensures that we are vague and don't assume anything in the prior.

So in summary

$$\begin{aligned}
\gamma_1 &\sim \mathrm{N}(-10, 100) \\
\gamma_2 &\sim \mathrm{N}(-5, 20) \\
\gamma_3 &\sim \mathrm{N}(0, 20) \\
\gamma_4 &\sim \mathrm{N}(5, 20) \\
\gamma_5 &\sim \mathrm{N}(10, 100) \\
\beta_1 &\sim \mathrm{N}(0, 20) \\
\alpha_1 &\sim \mathrm{N}(0, 10)
\end{aligned}$$

7.7.3 The rjags specification

We can now create our model file to compute our values using the following code

```
model

{
        for (i in 1:4) {
                y[i,]~dmulti(p[site[i],sex[i],],n[i])
                         }
                for (st in 1:2) {
                            for (sx in 1:2) {
                 p[st,sx,1]<-q[st,sx,1]
                 p[st,sx,2]<-q[st,sx,2]-q[st,sx,1]
                 p[st,sx,3]<-q[st,sx,3]-q[st,sx,2]
                 p[st,sx,4]<-q[st,sx,4]-q[st,sx,3]
                 p[st,sx,5]<-q[st,sx,5]-q[st,sx,4]
                 p[st,sx,6]<-1-q[st,sx,5]
                        for (h in 1:5) {
                                logit(q[st,sx,h])<-gamma[h]+alpha[sx]+beta[st]
                                        }
                                      }
                           }
        gamma0[1]~dnorm(-10,0.01)
        gamma0[2]~dnorm(-5,0.05)
        gamma0[3]~dnorm(0,0.05)
        gamma0[4]~dnorm(5,0.05)
        gamma0[5]~dnorm(10,0.01)
        gamma[1:5]<-sort(gamma0)
        alpha[1]~dnorm(0,0.1)
        alpha[2]<-0-alpha[1]
        beta[1]~dnorm(0,0.04)
        beta[2]<-0-beta[1]
}
```

7.8 Assessing the convergence

Using R, we find that the convergence criteria are met by this model and the trace plots in Figure 7.6 show the model is mixing and converging well and is sufficient to analyse.

7.9 Assessing the posterior distribution

From Figure 7.7 we can see that for each γ the variance of the posterior distribution is smaller than that of the prior distribution so we have learnt from our data and are now more certain about our results. The mean values have changed from -10,-5,0,5 and 10 to -8.954224, -3.52919, 0.5096215, 4.75139 and 8.735937 respectively.

We can see the effects of the sex of the centipede, α_1, specifically the male sex effect in Figure 7.8. From this plot we can see that the variance is smaller in the posterior distribution than in the prior distribution so we are more certain about our value of α_1. The mean has increased from 0 to 1.990766 which shows again that a male centipede is more likely to have fewer segments than a female centipede.

In Figure 7.9 we can see the distribution of β_1, the effect of embryonic development occurring in the wild. The variance of the posterior distribution is much smaller than the variance of the prior distribution so we are now more certain about the value of β_1.

Using R, we can say that the mean of β_1 has decreased from 0 to -0.1303214. In total, we have 2000 simulated values for β_1. We find the total number of values of $\beta_1 < 0$ is 108008. So we can say $\Pr(\beta_1 < 0) = 0.9004$. So although the change in mean is only small we can say that there is a high chance that there will be a little bit of an environmental effect occurring. So centipedes that undergo embryonic development in the wild are more likely to have more segments than a centipede whose embryonic development takes place in laboratory conditions at this particular site.

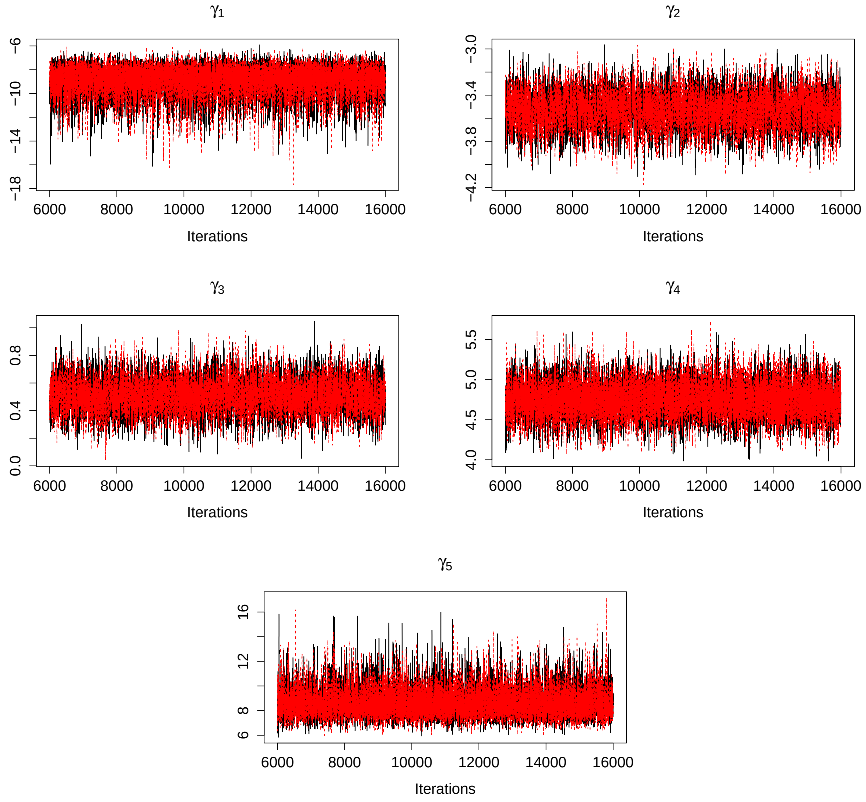

Figure 7.6: Trace plots of the γ's

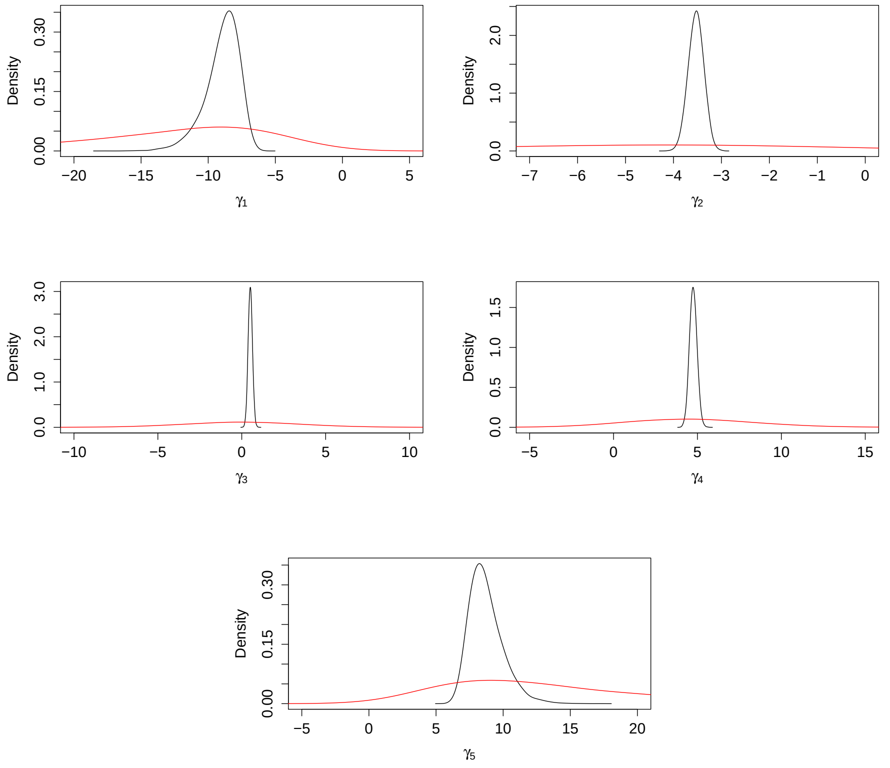

Figure 7.7: Prior (red) and posterior (black) density plots of the γ's

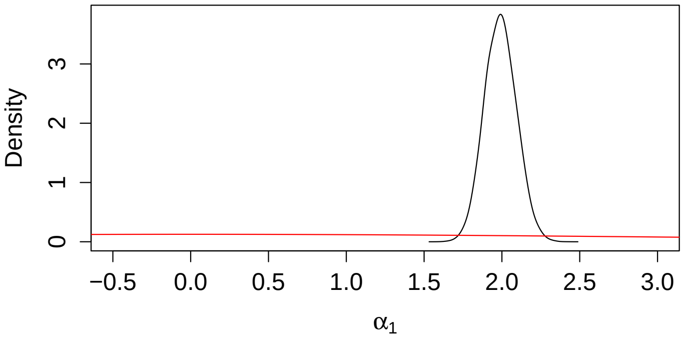

Figure 7.8: Prior (red) and posterior (black) density plot of α_1, the male sex effect

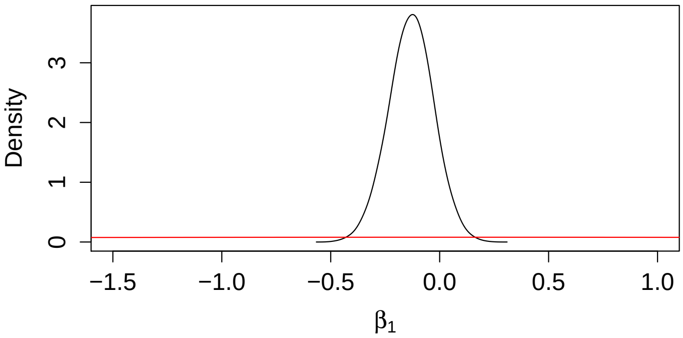

Figure 7.9: Prior (red) and posterior (black) density plot of β_1, the effect of the embryonic development occurring in the wild

Chapter 8

Conclusion

We have analysed a few factors that could have a potential effect on the segment number variation in centipedes. The first factor is environmental and the second genetic.

Performing Bayesian analysis on our latitudinal data in Chapter 4, we have reaffirmed the findings of Kettle and Arthur (2000) that a centipede found at a higher latitude is more likely to have fewer leg bearing segments than one found at a lower latitude and that females are, on average, bigger than males. After altering our model to include variance components in Chapters 5 and 6, we concluded that deviation from our line of regression is not just due to random sampling but also because of the individual variance found in samples from each latitude. Furthermore, we found that the sites that were taken close together, i.e. in a cluster, have a similar variation to each other and so should not be considered as independent within our analysis.

In Chapter 7, we analysed the data from Vedel et al. (2009) and ruled out the possibility that the segment number variation within centipedes is purely environmental and concluded that there is a genetic element affecting the development. We found that centipedes with a large number of segments generally produced larger offspring than a centipede with a smaller number of segments.

Comparing the data in Chapter 7 to that at the same site in Chapter 4, we found that there is also an environmental effect affecting the number of segments that a centipede develops. We saw that centipedes that undergo embryonic development in the wild have a slightly higher chance of developing more segments than a centipede undergoing embryonic development in laboratory conditions.

To investigate the segment number variation in centipedes further, we would look at other factors that could have an influence such as the study by Vedel et al. (2010) on the effects of the temperature during embryonic development.

Appendix

Model with addition of two random effects

```
model

{
        for (i in 1:26) {
                y[i,]~dmulti(p[site[i],sex[i],],n[i])
                          }

        for (st in 1:13) {
            for (sx in 1:2) {
                 p[st,sx,1]<-q[st,sx,1]
                 p[st,sx,2]<-q[st,sx,2]-q[st,sx,1]
                 p[st,sx,3]<-q[st,sx,3]-q[st,sx,2]
                 p[st,sx,4]<-q[st,sx,4]-q[st,sx,3]
                 p[st,sx,5]<-1-q[st,sx,4]
                        for (h in 1:4) {
                                logit(q[st,sx,h])<-gamma[h]+alpha[sx]
                                        +beta*(x[st]-55)+r[st]+w[cluster[st]]
                                        }
                              }
                            }
        for (s in 1:13) {
                r[s]~dnorm(0,tau.site)
                }
        for (cc in 1:10) {
                w[cc]~dnorm(0,tau.cluster)
                }
        tau.total~dgamma(1.1,0.3)
        delta~dbeta(1.5,1.5)
        tau.site<-tau.total/delta
        tau.cluster<-tau.total/(1-delta)
        gamma0[1]~dnorm(-10,0.05)
        gamma0[2]~dnorm(-5,0.05)
        gamma0[3]~dnorm(0,0.05)
        gamma0[4]~dnorm(5,0.05)
        gamma[1:4]<-sort(gamma0)
        alpha[1]~dnorm(0,0.1)
        alpha[2]<-0-alpha[1]
        beta~dnorm(0,0.04)
}
```

Bibliography

Atkinson, G., C. E. Taylor, and H. Jones (2010). Interindividual variability in the improvement of physiological risk factors for disease: gene polymorphisms or simply regression to the mean? *The Journal of Physiology 588*, 1023–1024.

Barnett, A. G., J. C. Pols, and A. J. Dobson (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology 34*, 215–220.

Brooks, S. P. (1998). Markov chain monte carlo methods and its application. *Journal of the Royal Statistical Society. Series D (The Statistician) 47*, 69–100.

Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. *The American Statistician 46*, 167–174.

Congdon, P. (2005). *Bayesian Models for Categorical Data.* Chichester : John Wiley & Sons, Ltd.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research 38*, 306–347.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association 95*, 1300–1304.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Gilks, W. R., A. Thomas, and D. J. Spiegelhalter (1994). A language and program for complex Bayesian modelling. *Journal of the Royal Statistical Society. Series D (The Statistician) 43*, 169–177.

Horneland, E. O. and B. A. Meidell (1986). The epimorphosis of strigamia maritima leach, 1817 chilopoda geophilidae. *Entomologica Scandinavica 17*, 127–129.

Kettle, C. and W. Arthur (2000). Latitudinal cline in segment number in an arthropod species, strigamia maritima. *Proceedings of the Royal Society of London, B 267*, 1393–1397.

Maher, M. and L. Mountain (2009). The sensitivity of estimates of regression to the mean. *Accident Analysis & Prevention 41*, 861–868.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological) 42*, 109–142.

Menard, S. (2002). *Applied Logistic Regression Analysis.* Sage University Papers Series on Quantitative Applications in the Social Sciences,series no. 07-106. Sage Publications.

O'Hagan, A., C. E. Buck, and A. Daneshkhah (2006). *Uncertain Judgements : Eliciting Experts' Probabilities.* Chichester : John Wiley & Sons, Ltd.

Plummer, M. (2012). *JAGS Version 3.3.0 user manual.*

Plummer, M. (2014). *Bayesian graphical models using MCMC.*

R Development Core Team (2004). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Smith, A. F. M. and A. E. Gelfand (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician 46*, 84–88.

Smith, A. F. M. and G. O. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological) 55*, 3–23.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–540.

Vedel, V., Z. Apostolou, W. Arthur, M. Akam, and C. Brena (2010). An early temperature-sensitive period for the plasticity of segment number in the centipede *Stigamia maritima. Evolution and Development 12*, 347–352.

Vedel, V., C. Brena, and W. Arthur (2009). Demonstration of a heritable component of the variation in segment number in the centipede strigamia maritima. *Evolution and Development 11*, 434–440.

Zellner, A. and C. Min (1995). Gibbs sampler convergence criteria. *Journal of the American Statistical Association 90*, 921–927.