Assessed practical 3: Investigating the effect of Easting and Northing on the number of beetle larvae in a maize field

Practical Number: P535

February 12, 2017

Abstract

A Bayesian approach is taken in order to infer on the parameters representing the effect of Easting and Northing on the number of Japanese beetles in a crop. A Monte Carlo Markov Chain (MCMC) is used to sample the posterior distributions of the parameters. It is argued that the MCMC run is of quality and that the parameters are well sampled. The Bayes factor between our two suggested normal priors is found to be 26 using the harmonic estimator of the marginal likelihood. The 95% higher posterior density confidence intervals are computed for each parameter as well as the posterior means. It is found that the Northing values do not have a big influence on the number of larvae, but that the Easting values largely contribute to the variation observed across the field.

Keywords: Bayesian inference; Japanese beetle; MCMC; Metropolis-Hastings; Poisson GLM;

1 Introduction

An 18×8 foot area of field planted with maize was split into 3×1 foot (0.28 square meter) rectangles (so 48 rectangles/plots in total). For each of the 48 plots, the number of Japanese beetle larvae found in the top foot of the soil was recorded. Table 1 below shows the recorded data. The Easting goes from A to H while the northing goes from 6 to 1.

		Easting							
		A	B	C	D	E	F	G	H
Northing	6	5	6	6	6	3	8	16	19
	5	9	8	5	3	6	3	1	11
	4	10	12	4	6	5	7	8	18
	3	6	6	10	8	6	4	10	10
	2	7	12	11	4	2	7	7	5
	1	11	12	8	7	3	6	3	10

Table 1: Contingency table for the number of beetle larvae for each of the 48 plots.

We are interested in assessing how the values of Easting and Northing impact the number of larvae on each plot. For instance Easting A could produce more larvae than Easting B. The two explanatory variables are then Easting and Northing and will be taken as categorical variables. The response is the number of larvae. A Bayesian analysis will be performed to answer the question.

Firstly, the data and the setting of the experiment will be quickly examined in order to fit an adequate likelihood and come up with reasonable priors for the parameters of interest. After which, the posterior will be sampled using a Monte Carlo Markov Chain (MCMC) and a few diagnostics will be performed on the MCMC run to assess its quality. Finally, we will compute the Bayes Factor to compare our different priors, and inference on the parameters will be made on the model with the most adequate prior. The statistical analysis is done using the software R and the code is provided in Appendix A.

2 Preliminary analysis

2.1 Likelihood

Figure 1 below displays Table 1 in a more visual way to see how Northing and Easting may affect the distribution of larvae.

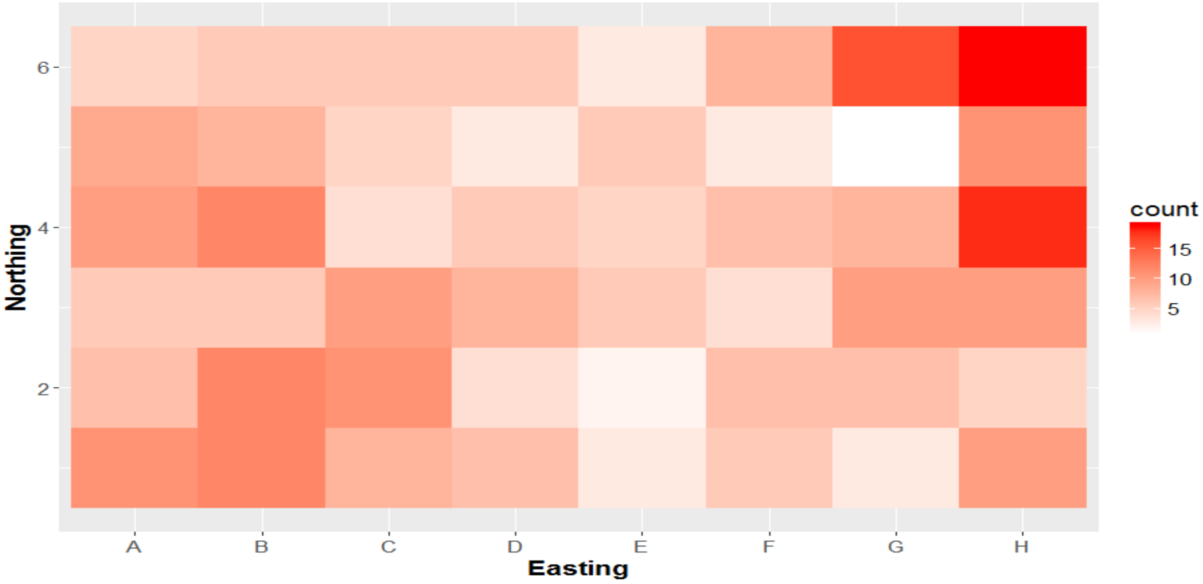

Figure 1: Heatmap for the observed data in Table 1.

It is not obvious that any of the Northing values increase the number of larvae. However it seems that Easting H and B give rise to a higher number of larvae. Since the response of interest is a count data, the likelihood will be taken as a Poisson GLM with 14 parameters (6 for Northing and 8 for Easting). We will not use interaction terms as it would give too many parameters. Therefore the likelihood is

$$f(y|\beta) = \prod_{i=1}^{48} \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \propto \prod_{i=1}^{48} \exp(-\mu_i)\mu_i^{y_i}, \tag{1}$$

where

$$\mu_i = \exp(x_i\beta). \tag{2}$$

In equation (2), x_i is the indicator vector corresponding to variable y_i. The y_i variables are taken from the matrix in Table 1 and put into a vector column by column.

The column vector β is a vector of 14 parameters as mentioned previously and $\beta = (\beta_{Northing6}, ..., \beta_{Northing1}, \beta_{EastingA}, ..., \beta_{EastingH})^T$. Therefore, x_1 would have a 1 as the first and seventh elements and zero for the remaining twelve elements.

2.2 Prior elicitation

We want to sample the posterior $p(\beta|y)$, where as usual

$$p(\beta|y) \propto f(y|\beta)\pi(\beta). \tag{3}$$

We want to find a sensible prior $\pi(\beta)$. First of all, we will assume that a uniform number of larvae across all of the 48 plots is a sensible thing. We have no prior knowledge leading us to think otherwise. Thus we can take that the parameters are independent and so the prior can just factorize into the product of the 14 priors for each parameter. Since we want the number of larvae to be uniform we will take each parameter to have the same distribution. We chose the prior for each parameter to be a normal distribution. The posterior is then

$$p(\beta|y) \propto \prod_{i=1}^{48} \exp(-\mu_i)\mu_i^{y_i} \prod_{j=1}^{12} \pi_j(\beta), \tag{4}$$

where $\pi_j(\beta) \sim N(0, \sigma^2)$. Figure 2 belows shows histograms for the simulated number of larvae using 2 different normal priors. The histogram on the left is for $N(0, 1)$ and the histogram on the right $N(0, 5)$.

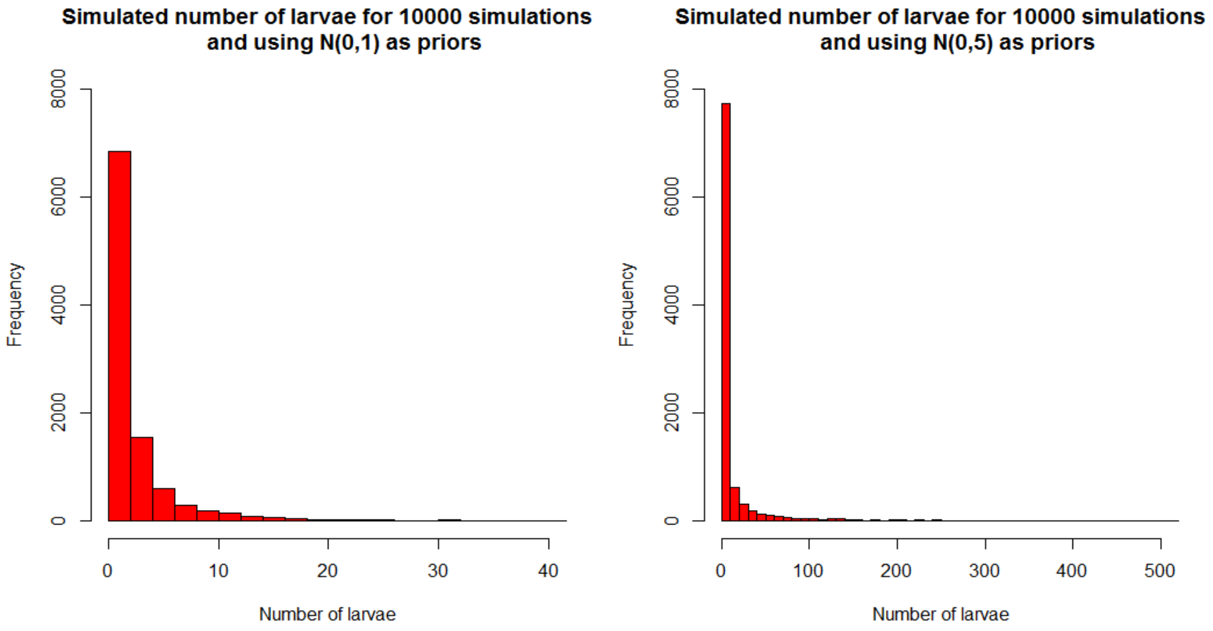

Figure 2: Histograms for the simulated number of larvae under two different priors.

One can see that both priors are heavily biased in favor of a small number of larvae. This is fine as each plot of maize are about 0.27 square meter and one would not expect a very high number of larvae in such a small area. However, since we are not sure how the number of larvae may vary, we also try a prior with a larger variance. The second prior gives instances of much higher (of order 10 times higher) number of larvae. Let us now look into the MCMC and assess which prior is preferred.

3 MCMC

We ran an MCMC using the Metropolis-Hastings algorithm for 100000 samples. The proposal distribution is the same for all the parameters, namely a $N(0, 0.1)$. The size of the variance was tuned such that we get a reasonable acceptance rate. Figure 3 below shows the time series and histogram for the parameter $\beta_{Northing6}$. We could also show similar plots for all of the 14 parameters but this is easily reproducible using the code in Appendix A.

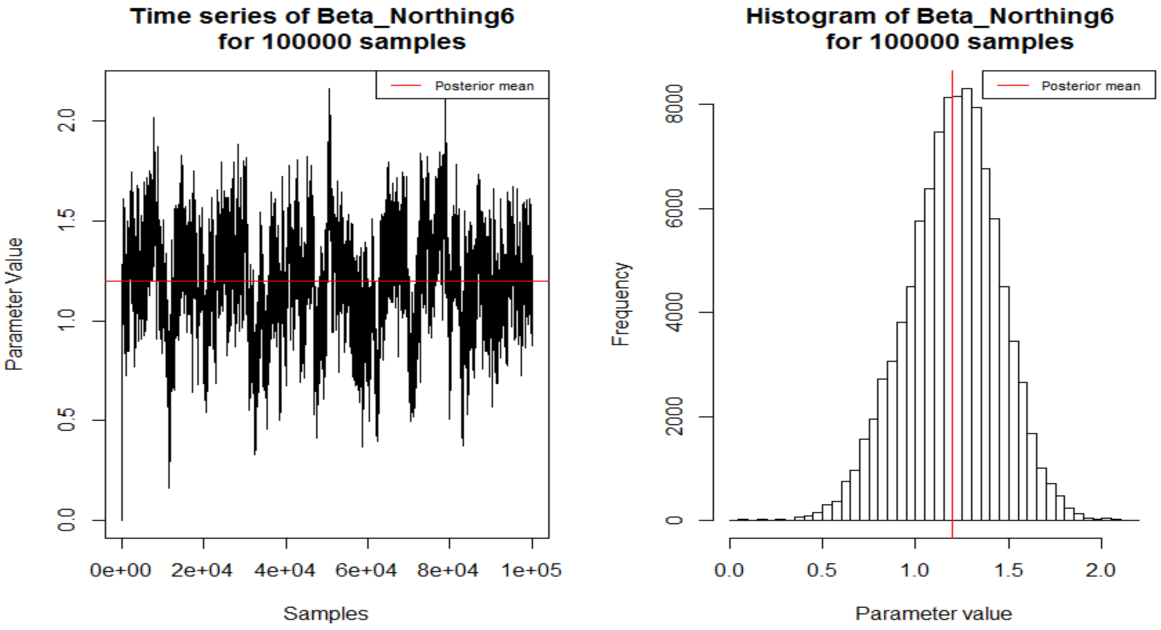

Figure 3: Time series and histogram for $\beta_{Northing6}$ with the posterior mean represented as a red line.

One can see from the plot on the left that there is a short burn-in period of a few hundred samples but the chain quickly converges and starts sampling around the same value. The histogram is smooth and shows that the parameter is well sampled. Note that since the chain was run for 100000 iterations, the burn-in phase is not removed to compute the posterior mean since it has close to no effect. In addition, the quality of the MCMC run will be assessed in the two following ways. The posterior distribution for several runs with different starting values is plotted in Figure 4, alongside the autocorrelation of the first $\beta_{Northing6}$ samples and of the log-posterior.

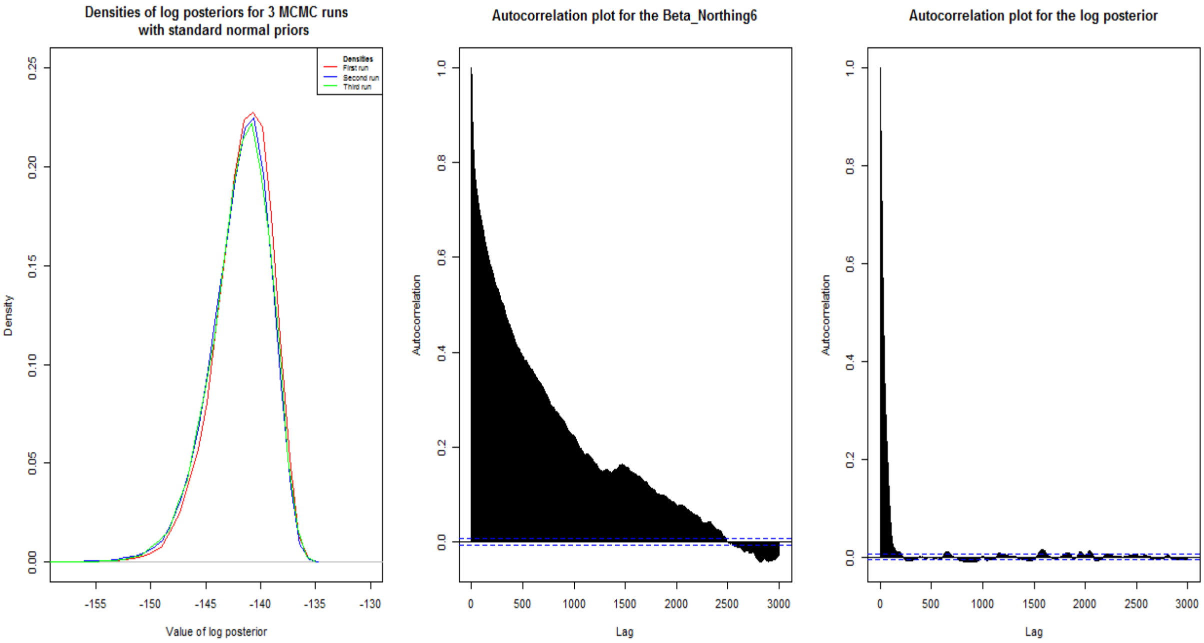

Figure 4: Densities of the log posterior for 3 different runs and autocorrelation plots for $\beta_{Northing6}$ samples and the log-posterior for the first MCMC run.

The left hand side plot in Figure 4 shows that the three different MCMC runs with different starting values give rise to a very similar log posterior distribution. The plot in the middle shows the autocorrelation for $\beta_{Northing6}$. One can see that it takes a large number of samples for the autocorrelation to go down to 0, but it eventually attains it at a lag of 2500. Then, if we wanted to have a sample of uncorrelated values for $\beta_{Northing6}$, we would take every multiple of $2500th$ and discard the remaining samples. It would mean that one would have to run the MCMC for an extremely large number of iterations in order to have a large uncorrelated sample. We have seen that the chain converges, multiple runs give rise to the approximately the same log posterior distribution and that the autocorrelation between samples goes down to 0. Therefore, the MCMC run is of quality and we can make inference on the parameters of interest. However, before computing the posterior means and higher posterior density intervals for the parameters, we need to know which prior is favored, by computing the Bayes factor.

4 Bayes factor and inference

4.1 Bayes factor

The Bayes factor is computed in order to have a quantitative assessment on which of $N(0,1)$ and $N(0,5)$ is a better prior, if any. Let m_1 denote the model for the first prior (standard normal) and m_2 for the second prior. The Bayes factor B is the ratio of the marginal likelihoods,

$$B = \frac{p(y|m_1)}{p(y|m_2)} = \frac{\int_\beta f(y|\beta)\pi_{m_1}(\beta)d\beta}{\int_\beta f(y|\beta)\pi_{m_2}(\beta)d\beta}. \tag{5}$$

Here, we use the harmonic estimate of the marginal likelihood and the estimated value for the Bayes factor is 26. This is a very high value and thus we have strong evidence in favor of the standard normal prior. We are now in a position where inference on the parameters can be made using the first prior.

4.2 Inference on the parameters

The question of interest was to investigate the effect of the Northing and Easting on the number of larvae. The point estimates (posterior means) and highest posterior density intervals are recorded in Table 2 below.

Parameters	Estimates	95% HPD CI	Parameters	Estimates	95% HPD CI
$\beta_{Northing6}$	1.26	[0.70 1.66]	$\beta_{EastingB}$	1.08	[0.65 1.65]
$\beta_{Northing5}$	0.86	[0.25 1.27]	$\beta_{EastingC}$	0.84	[0.39 1.44]
$\beta_{Northing4}$	1.27	[0.73 1.69]	$\beta_{EastingD}$	0.57	[0.13 1.19]
$\beta_{Northing3}$	1.11	[0.53 1.52]	$\beta_{EastingE}$	0.26	[-0.22 0.92]
$\beta_{Northing2}$	1.04	[0.47 1.45]	$\beta_{EastingF}$	0.60	[0.11 1.19]
$\beta_{Northing1}$	1.12	[0.55 1.55]	$\beta_{EastingG}$	0.86	[0.42 1.44]
$\beta_{EastingA}$	0.92	[0.48 1.51]	$\beta_{EastingH}$	1.35	[0.96 1.91]

Table 2: Posterior means of the parameters with their associated HPD 95% confidence intervals.

We can see that the Northing categories do not seem to give rise to significant differences in the predicted number of larvae as the posterior means are all quite close to each other. However, it is worth noting that Northing 5 is somewhat lower than other Northing values (mostly due to the very low number of larvae on Easting G). On the other hand, the Easting values seem to exhibit more differences. Indeed, A, B, G and H all have parameters with high posterior mean, whereas D, E and F have considerably smaller posterior means. That is even with a prior that led to uniform number of larvae across the field (which is not the case when one looks at the collected data), the posterior still displays the pattern that is observed in the data.

5 Conclusions

We used a Poisson GLM to fit the count data from Table 1 and then performed a prior elicitation which led to using normal priors on the parameters. We argued that the MCMC run was of quality and found that the computed Bayes factor heavily favored the standard normal against our second prior $N(0,5)$. The posterior means were computed for each of the 14 parameters and were used in order to answer the question of interest. The number of larvae do not seem to vary significantly across the Northing values. However, Easting A, B, G and H all exhibit a significantly (since our prior was conservative and giving a uniform number of larvae for each plot) higher number of larvae, whereas D, E and F are considerably smaller. This gives rise to a quadratic pattern across the Easting values. Using the posterior means of the parameters to predict the number of larvae in each plot leads to poor result as the predictions tends to be too small. This is due to a very strong bias towards a number of larvae close to 0 from our prior, as was highlighted in Figure 2. It would thus be of interest in a subsequent experiment to use our current posterior knowledge as a prior to get better predictions.

Appendix A

```
larva<- read.table(file="C:/Users/pinouche/Downloads/beetlelarva.txt",header=TRUE)
fix(larva)



library(lattice)
library(nlme)
library(MASS)
install.packages("Rcpp")
install.packages("ggplot2")
library(Rcpp)
library(grid)
library(ggplot2)
require(reshape2)
require(ggplot2)
install.packages(’sna’)
library(’sna’)
install.packages(’coda’)
library(coda)

# explanatory analysis with a heatmap of the data

Easting <- larva$easting
Northing <- larva$northing

ggplot(larva, aes(Easting, Northing)) +
  theme(plot.title = element_text(face="bold",
                                  size=20, hjust=0.5)) +
  theme(axis.title = element_text(face="bold", size=16,
                                  hjust=0.5)) +
  geom_tile(aes(fill = count)) +
  scale_fill_gradient(low = "white", high = "red") +
  theme(text = element_text(size=16))

# Prior elicitation

# Prior number 1

param1 <- rnorm(10000, mean = 0, sd = 1)
param2 <- rnorm(10000, mean = 0, sd = 1)

expvector1 <- 0
for (i in 1:10000)
{
  expvector1[i] <- exp(param1[i]+param2[i])
}

# Priot number 2

param1 <- rnorm(10000, mean = 0, sd = sqrt(5))
param2 <- rnorm(10000, mean = 0, sd = sqrt(5))

expvector <- 0
for (i in 1:10000)
{
  expvector[i] <- exp(param1[i]+param2[i])
}

# Plot histograms of the simulated data

par(mfrow = c(1,2))
hist(expvector1,breaks=200,col=’red’,xlim=c(0,40),ylim=c(0,8000),xlab = ’Number of larvae’,main=’Simulated number of larvae for 10000 simulations
      and using N(0,1) as priors’)
hist(expvector,breaks=10000,col=2,xlim=c(0,500),ylim=c(0,8000),xlab = ’Number of larvae’,main=’Simulated number of larvae for 10000 simulations
      and using N(0,5) as priors’)

# Create the X matrix

to_dummy = function(X) {
  out = data.frame(matrix(nrow=nrow(X), ncol=0))
  for (val in unique(X$northing)) {
    out[paste("northing", val, sep="_")] = ifelse(X$northing==val, 1, 0)
  }

  for (val in unique(X$easting)) {
    out[paste("easting", val, sep="_")] = ifelse(X$easting==val, 1, 0)
  }

  return(data.matrix(out))
}
mat <- to_dummy(larva)

# likelihood


loglik <- function(larva,beta)
{
  loglik <- 0
  mu <- 0

  for (i in 1:48)
  {

    mu[i] <- exp(mat[i,]%*%beta)
    loglik <- loglik + log(dpois(larva$count[i],mu[i]))

  }
  return(loglik)
}

# First Prior

lpr<-function(beta) {
  sum(log(dnorm(beta)))
}


# Second Prior

lpr<-function(beta) {
  sum(log(dnorm(beta,sqrt(10))))
}

#initialise (could use glm fit)

beta0=c(rep(0,times=14))

#MCMC loop - here "beta" is the current state of the Markov chain.
#betap will be the proposed state

MCMC<-function(K=100000,beta=beta0) {
  #returns K samples from posterior using MCMC
  #no subsampling, start state goes in beta

  B=matrix(NA,K,14); LP=rep(NA,K); LLK=rep(NA,K)
  #storage, I will write the sampled betas here

  lp=loglik(larva,beta)+lpr(beta)
  #log posterior is log likelihood + log prior + constant

  for (k in 1:K) {

    #tuned RW MH - I adjusted the step sizes so they were
    #unequal for beta[1] and beta[2]
    betap=rnorm(14,beta,0.1)#generate candidate

    LLK[k]=loglik(larva,betap)
    lpp=LLK[k]+lpr(betap)        #calculate log post for candidate

    MHR=lpp-lp                         #"log(pi(y)/pi(x))"
    print(MHR)
    if (log(runif(1))<MHR)
      {           #Metropolis Hastings acceptance step
```

```r
      beta=betap                       #if we accept update the state
      lp=lpp
    }

    B[k,]=beta                         #save the sequence of MCMC states, our samples.
    LP[k]=lp
  }
  return(list(B=B,L=LP,LL=LLK))
}

# Stored values for 2 runs with prior 1
K <- 100000
Output1 <- MCMC(K,beta=beta0);
beta1<- c(rep(1,times=14))
beta2 <- c(rep(10,times=14))
Output2 <- MCMC(K,beta=beta1);
Output3 <- MCMC(K,beta=beta2);

Run1Prior1B <- Output1$B # Result of the second MCMC run with different starting values
Run1Prior1LP <- Output1$L
Run1Prior1LLLK1 <- Output1$LL
Run2Prior1B <- Output2$B # Result of the second MCMC run with different starting values
Run2Prior1LP <- Output2$L
Run2Prior1LLLK1 <- Output2$LL
Run3Prior1B <- Output3$B # Result of the second MCMC run with different starting values
Run3Prior1LP <- Output3$L
Run3Prior1LLLK1 <- Output3$LL

# Plot the densities for each run on top of each other

par(mfrow=c(1,3))
plot(density(Run1Prior1LP), xlim=c(-158,-130),col='red',ylim=c(0,0.25),xlab='Value of log posterior',
     main = 'Densities of log posteriors for 3 MCMC runs
    with standard normal priors')
lines(density(Run2Prior1LP),col='blue')
lines(density(Run3Prior1LP),col='green')
legend( 'topright', inset=0,
        legend=c(expression(bold("Densities")),"First run","Second run",'Third run'),
        col=c(NA,'red','blue','green'),
        lty=c(NA,1,1,1), merge=FALSE, cex=0.7)

acf(Run1Prior1B[,1],lag.max=3000,ylab='Autocorrelation',main='Autocorrelation plot for the Beta_Northing6')
acf(Run1Prior1LP,lag.max=3000,ylab='Autocorrelation',main='Autocorrelation plot for the log posterior')


# convergence plot for Beta_Northing6

plot(Run1Prior1B[,1],type='l',xlab='Samples',ylab='Parameter Value',main='Time series of Beta_Northing6
     for 100000 samples')
abline(h=mean(Run1Prior1B[,1]),col='red')
legend( 'topright', inset=0,
        legend=c('Posterior mean'),
        col=c('red'),
        lty=c(1), merge=FALSE, cex=0.7)

hist(Run1Prior1B[,1],breaks=50,xlab='Parameter value',ylab='Frequency',main='Histogram of Beta_Northing6
     for 100000 samples')
abline(v=mean(Run1Prior1B[,1]),col='red')
legend( 'topright', inset=0,
        legend=c('Posterior mean'),
        col=c('red'),
        lty=c(1), merge=FALSE, cex=0.7)
# Stored values for 1 run with the second prior N(0,10)

Run1Prior2B <- Output1$B # Result of the second MCMC run with different starting values
Run1Prior2LP <- Output1$L
Run1Prior2LLLK1 <- Output1$LL

# Plot the 2 densities for the 2 runs to see if they agree (first prior)

# Computing the harmonic estimates for the two priors and then compute bayes factor
p_hat <- 1/(mean(1/(exp(Run1Prior1LLLK1))))
p_hat2 <- 1/(mean(1/(exp(Run1Prior2LLLK1))))

# Bayes factor
Bayesfac <- (p_hat/p_hat2)

# Get the HPD 95% confidence itnerval

HPDinterval(as.mcmc(Run1Prior1B))

# Obtain the posterior mean for each parameter

PosteriorMeanVec <- 0
for (i in 1:14)
{

 PosteriorMeanVec[i] <- mean(BB[,i])

}

PosteriorMeanVec

# Obtain the predicted number of larvas on each of the 48 plots, using the posterior means as point estimates

mu2 <- 0

for (i in 1:48)
{

  mu2[i] <- exp(mat[i,]%*%PosteriorMeanVec)

}
```