A Simulation Study to Investigate the Bayesian Non-linear Regression Model of Daily Pollen Counts across One Year

1. Introduction

Pollen is a powdery substance produced by plants for propagation. The pollen counts follow a seasonal pattern because plant reproduction favors a particular period of time, mostly spring, during a year. In the study, several values of sample size n and correlation parameter ρ are selected and for each combination of n and ρ, 100 datasets are generated. The goal is to test the performance of the Bayesian non-linear regression model of daily pollen counts given the simulated datasets.

2. Model

The response variable, Y_t, represents the measured daily pollen counts on day t where t ranges from 1 to 365. Since pollen counts have a seasonal pattern, the following non-linear regression model is considered:

$$Y_t = \mu_t + \varepsilon_t$$

where

$$\mu_t = b_1 + b_2 \exp\left\{\frac{-(t-b_3)^2}{2b_4^2}\right\}$$

and

$$\varepsilon_t \sim Normal(0, \sigma^2), Cor(\varepsilon_t, \varepsilon_{t+h}) = \rho^h$$

We plot several simulated datasets and observe that the distribution of Y_t is a bell curve. Meanwhile, the formula of μ_t is very similar to the probability density function of normal distribution. Thus, Y_t is likely to follow a normal distribution. Assuming errors are independent across time (i.e., $\rho = 0$) when fitting the model, Y_t is modeled using normal likelihood $Y_t \sim Normal(\mu_t, \sigma^2)$ and the uninformative priors are shown below.

$$b_1, b_2, b_4 \sim Normal(0, 0.0001), b_3 \sim Uniform(1, 365), \tau^2 \sim Gamma(0.1, 0.1)$$

To make sure the uninformative prior of b_1, b_2, and b_4 covers the parameter's support, the corresponding value of precision is set to very small.

3. Computation

To analyze a simulated dataset, an R function named `pollen_fit` is written. The function takes vector t and Y as inputs and returns the posterior mean and 95% credible interval for each b_j where j = 1, 2, 3, 4. Within the function, the model is constructed in R using the package `rjags`. The code is attached in the appendix. The specifications for building the model are 2 chains, 5000 burn-in, 10000 iterations, and no thining. For each combination of n and ρ, three datasets are randomly selected to check convergence. Based on the trace plots, the model converges very well. Below is a trace plot of b_2 given one of the simulated datasets randomly selected for the combination of n = 100 and ρ = 0.9. The effective sample size of each parameter is much larger than 1000 and all the values of Gelman Rubin statistics are equal to 1. Posterior predictive checks are also conducted on the selected simulated datasets. Given the same simulated dataset that is used to generate the trace plot of b_2, the Bayesian p-values computed for minimum of Y, maximum of Y, range of Y, and mean of Y are respectively 0.355, 0.186, 0.258, and 0.500. Thus, the posterior predictive distribution of Y matches the simulated data well. Based on the distribution of several simulated datasets and the representation of each b_j, initial values of b_j are chosen using formulas shown in Table 1. For the initial value of b_1, median of Y is better to capture the baseline mean than the mean of Y when there is a group of data points that significantly deviate from the baseline mean. For the initial value of b_2, the

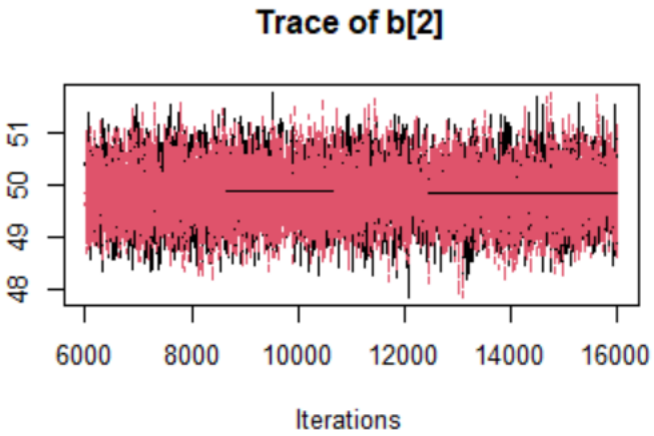

Figure 1: The Trace Plot of b_2 given #24 simulated dataset for the combination of n=100 and ρ=0.9

	Parameter Representation	Initial Value Formula in R Code
b_1	the baseline mean	median(Y)
b_2	the increase in the mean at the peak of pollen season	max(Y) – median(Y)
b_3	the day of the year with the highest mean	t[which.max(Y)]
b_4	the width of the peak	sd(Y)

Table 1: The Representation and Formula of Selecting Initial Values in R Code for Each Parameter b_j

maximum of Y subtracted by the baseline mean captures the increase in the mean at the peak of pollen season. For the initial value of b_3, it is simply the value of t that maximizes Y. And for the initial value of b_4, the standard deviation of Y captures the width of the peak since the variability of Y is mainly caused by the data points in pollen season diverging from the baseline mean.

4. Data Generation

The combinations of n and ρ explored in the study are present in Table 2. For each combination of n and ρ, 100 datasets are generated using the code provided in the final exam instruction within a for loop. The same values are used for all the other parameters in the code. Within each for loop, a dataset is generated based on its unique seed number, which ensures the same results when executing the code as well as makes it easier for others to reproduce the results.

Combination	Sample Size n	Correlation Parameter ρ
1	100	0.9
2	100	0.5
3	200	0.9
4	200	0.5

Table 2: The Combinations of Sample Size n and Correlation Parameter ρ

5. Metrics

For each combination of n and ρ, the metrics utilized to summarize the performance of the Bayesian non-linear regression model are the bias, mean squared error, and coverage of 95% credible intervals for each of the four b_j. The Monte Carlo standard errors of the bias, mean square error, and coverage are also reported. The formula of each metric is shown in Table 3. The posterior mean of b_j for dataset $s \in \{1, \ldots, 100\}$ is $\widehat{b_j^s}$ and the true value of b_j is b_j^*. s_j is the standard deviation of $\{\widehat{b_j^1}, \ldots, \widehat{b_j^{100}}\}$. $l\left(\widehat{b_j^s}\right)$ and $u\left(\widehat{b_j^s}\right)$ respectively represent the lower bound and the upper bound of the 95% credible interval for $\widehat{b_j^s}$. And $P\left(l\left(\widehat{b_j^s}\right) < b_j^* < u(\widehat{b_j^s})\right)$ is the possibility of b_j^* included in the

Metrics	Formula	Monte Carlo SE	Formula
Bias	$\sum_{s=1}^{100}(\widehat{b_j^s} - b_j^*)/100$	**Bias SE**	$s_j/\sqrt{100}$
MSE	$\sum_{s=1}^{100}\left(\widehat{b_j^s} - b_j^*\right)^2/100$	**MSE SE**	$s_j/\sqrt{100}$
Coverage	$\sum_{s=1}^{100} P\left(l\left(\widehat{b_j^s}\right) < b_j^* < u(\widehat{b_j^s})\right)/100$	**Coverage SE**	$s_j/\sqrt{100}$

Table 3: The Formulas and Corresponding Monte Carlo Standard Errors for Each Metrics

95% credible interval. Note that the true value b_j^* is either included in or excluded from the 95% credible interval. Hence the value of $P\left(l\left(\widehat{b_j^s}\right) < b_j^* < u(\widehat{b_j^s})\right)$ is either 0 or 1 for each $\widehat{b_j^s}$.

6. Results

For each combination, the posterior mean and 95% credible interval for each b_j is shown in Table 4. As mentioned in the exam, the true values of b_1, b_2, b_3, and b_4 are 10, 50, 100, and 10, respectively. For all four combinations, the posterior mean of each parameter is very close to the true value. For combinations 1 and 2, all the 95% credible intervals include the true value. However, the 95% credible intervals of b_1 and b_3 exclude the true values for combination 3, and the 95% credible intervals of b_2 and b_4 exclude the true values for combination 4. Thus, the 95% credible intervals of b_j for combinations with smaller sample size n turn out to be more accurate than the 95% credible intervals of b_j for combinations with larger sample size n. For each combination, the bias, MSE, coverage of 95% credible interval, and their corresponding Monte Carlo standard errors for each parameter are present in Table 5. The values of bias and MSE are very small for each b_j in all four combinations but most values of coverage are much lower than 0.95. The Monte Carlo standard errors of each metric are very small. Comparing between combination 1

Combination	Parameter	Posterior Mean	95% Credible Interval
Combination 1: n = 100, ρ = 0.9	b_1	9.881044	(9.688128, 10.108260)
	b_2	49.769529	(49.69306, 51.92431)
	b_3	99.946593	(99.91532, 100.40162)
	b_4	10.142225	(9.545696, 10.060485)
Combination 2: n = 100, ρ = 0.5	b_1	9.899483	(9.688128, 10.108260)
	b_2	50.807041	(49.69306, 51.92431)
	b_3	100.160061	(99.91532, 100.40162)
	b_4	9.800313	(9.545696, 10.060485)
Combination 3: n = 200, ρ = 0.9	b_1	10.179611	(10.04854, 10.30952)
	b_2	49.496588	(48.82326, 50.17856)
	b_3	99.709156	(99.55500, 99.86532)
	b_4	9.848485	(9.687345, 10.007420)
Combination 4: n = 200, ρ = 0.5	b_1	10.010350	(9.852564, 10.166422)
	b_2	51.057559	(50.24058, 51.87493)
	b_3	99.943509	(99.76497, 100.12459)
	b_4	9.811207	(9.629939, 9.992370)

Table 4: The Posterior Mean and 95% Credible Interval for Each b_j For Each Combination

and 2, we see that larger ρ leads to smaller coverage. The comparison between combination 3 and 4 shows the same result. Comparing between combination 1 and 3, we see that smaller n leads to larger coverage. The comparison between combination 2 and 4 shows the same result. For each parameter, the bias, MSE, and the Monte Carlo standard errors of bias, MSE, and coverage are nearly the same for all four combinations.

Combination	Parameter	Bias	MSE	Cov	Bias SE	MSE SE	Cov SE
Combination 1: n = 100, ρ = 0.9	b_1	-0.00921	0.07319	0.58	0.02717	0.01094	0.04960
	b_2	0.01277	0.60008	0.81	0.07784	0.08787	0.03943
	b_3	-0.00673	0.04139	0.75	0.02044	0.00537	0.04352
	b_4	0.03165	0.03912	0.76	0.01962	0.00596	0.04292
Combination 2: n = 100, ρ = 0.5	b_1	0.00627	0.02037	0.90	0.01433	0.00254	0.03015
	b_2	-0.06228	0.30934	0.96	0.05555	0.03869	0.01969
	b_3	-0.00178	0.02142	0.91	0.01471	0.00299	0.02876
	b_4	0.00179	0.02220	0.89	0.01497	0.00294	0.03145
Combination 3: n = 200, ρ = 0.9	b_1	0.02232	0.06511	0.41	0.02555	0.00873	0.04943
	b_2	-0.02750	1.00948	0.40	0.10094	0.10111	0.04924
	b_3	-0.01232	0.04304	0.56	0.02081	0.00544	0.04989
	b_4	0.02119	0.03756	0.60	0.01936	0.00532	0.04924
Combination 4: n = 200, ρ = 0.5	b_1	0.00112	0.00840	0.88	0.00921	0.00117	0.03266
	b_2	-0.06603	0.26801	0.90	0.05161	0.03920	0.03015
	b_3	-0.00210	0.01346	0.89	0.01166	0.00177	0.03145
	b_4	0.01338	0.01639	0.78	0.01280	0.00199	0.04163

Table 5: The Metrics and the Monte Carlo Standard Errors of the Metrics for Each Parameter in Each Combination

7. Discussion

Based on the results, Bayesian non-linear regression is appropriate for modeling the daily pollen counts across one year when the sample size is 100 and the correlation of errors is 0.5. When the sample size increases to 200 and the correlation of errors increases to 0.9, Bayesian non-linear regression is not ideal anymore for modeling the daily pollen counts due to low coverage of 95% credible interval for each parameter b_j. Thus, it is appropriate to use Bayesian non-linear regression to analyze the real pollen count data when the sample size is relatively small and the errors are nearly independent. Bayesian non-linear regression is also useful to estimate each parameter in the model if the sample size is large and the errors are correlated but not recommended because of the low coverage of 95% credible intervals.

Appendix

```
##ST440 Final Exam##
##Sibo Peng##

# function for computation
pollen_fit <- function(t,Y){
  library(rjags)

  model_string <- textConnection("model{
    # likelihood
    for(i in 1:n){
      Y[i] ~ dnorm(mu[i], taue)
      mu[i] = b[1] + exp(-(t[i]-b[3])^2 / (2*(b[4])^2))*b[2]
    }

    # priors
    b[1] ~ dnorm(0,0.0001)
    b[2] ~ dnorm(0,0.0001)
    b[3] ~ dunif(1,365)
    b[4] ~ dnorm(0,0.0001)
    taue ~ dgamma(0.1,0.1)
  }")

  burn <- 5000
  iters <- 10000
  chains <- 2

  n <- length(Y)
  b1 <- median(Y)
  b2 <- max(Y)-median(Y)
```

```
  b3 <- t[which.max(Y)]

  b4 <- sd(Y)

  init_b <- c(b1, b2, b3, b4)


  data    <-list(Y=Y, t=t, n=n)

  inits   <-list(b=init_b)

  model   <-jags.model(model_string, data=data, init=inits,
n.chains=chains, quiet=TRUE)

  update(model, burn, progress.bar="none")

  samples <-coda.samples(model, variable.names=c("b"), n.iter=iters,
progress.bar="none")


  b1 <- unlist(samples[,1])

  b2 <- unlist(samples[,2])

  b3 <- unlist(samples[,3])

  b4 <- unlist(samples[,4])


  b <- cbind(b1, b2, b3, b4)

  out <- list("mean"=colMeans(b), "CI"=apply(b, 2, quantile,
(c(0.025,0.975))))


  output <- list(post_mean=out[[1]], cred_interval=out[[2]])

return(output)}
```