**STAT 451: Computational Statistics**

**Final Project**

**Statistical Evaluation of Hormone Treatment on Breast Cancer**

**Supervised By: Dr. Jie Yang**

**Team Members:**

|  |  |
| --- | --- |
| **Name** | **UIN** |
| Yue Zhang |  |
| Xiaohan Liu | 659692941 |
| Yang Chen | 668326530 |

**Table of Contents**

[1. Project Summary 4](#_Toc480968168)

[***1.1 Problem Description*** 4](#_Toc480968169)

[***1.2 Statistics Evaluation*** 4](#_Toc480968170)

[***1.3 Major Conclusions*** 4](#_Toc480968171)

[2. Technic Report 5](#_Toc480968172)

[***2.1 Basic Statistics of Breast Cancer Data*** 5](#_Toc480968173)

[***2.2 Parameter Estimation of Posterior Distribution*** 6](#_Toc480968174)

[***2.3 Convergence Diagnostics of Sampling Process*** 7](#_Toc480968175)

[***2.4 Hypothesis Test of Hormone Treatment*** 8](#_Toc480968176)

[***2.5 Sensitivity Analysis of Estimation*** 9](#_Toc480968177)

[3. Code in R 10](#_Toc480968178)

[***3.2 Gibbs Sampler*** 10](#_Toc480968179)

[***3.3 Sampling Convergence*** 11](#_Toc480968180)

[4. References 12](#_Toc480968181)

**List of Figures**

[Figure 1. Boxplots of recurrence 5](#_Toc480968182)

[Figure 2. Boxplots of censoring 5](#_Toc480968183)

[Figure 3. Survival function constructed by Kaplan Meier estimator 6](#_Toc480968184)

[Figure 4. Sample path for , and 8](#_Toc480968185)

[Figure 5. Cusum plot for , and 8](#_Toc480968186)

[Figure 6. Autocorrelation plot for , and 8](#_Toc480968187)

[Figure 7. Prior (curve) and estimated posterior distribution (histogram) for 9](#_Toc480968188)

**List of Tables**

[Table 1. Observed data for breast cancer treatment 5](#_Toc480968189)

[Table 2. Basic statistics of recurrence 5](#_Toc480968190)

[Table 3. Basic statistics of censoring 5](#_Toc480968191)

[Table 4. Parameter estimation results 9](#_Toc480968192)

[Table 5. Parameter estimation results of under different hyper-parameter settings 10](#_Toc480968193)

[Table 6. Parameter estimation results of under different hyper-parameter settings 10](#_Toc480968194)

# 1. Project Summary

## ***1.1 Problem Description***

In this project, whether hormone treatment could benefit women for breast cancer treatment is investigated by comparing with traditional cancer control methods.

The experiment is designed as follows: each subject entered the clinical trial when she had a recurrence. She was then treated by irradiation and assigned to either a hormone therapy group or a control group. The observation of interest is the time until a second recurrence, which is assumed as exponential distribution for both hormone therapy group (parameter ) and control group (). Many of the women did not have a second recurrence before the clinical trial was concluded, so that their recurrence times are censored. A censor time *M* means that a woman was observed for *M* months and did not have a recurrence during that period, so, her recurrence time is known to exceed *M* months.

Totally, 15 women who received the hormone treatment suffered recurrences, and their recurrence times is 280 months while 10 women who are in control group suffered recurrences with 189 times. Detail data of the experiments is included in Section 2.1.

## ***1.2 Statistics Evaluation***

To evaluate the effectiveness of hormone treatment, Bayesian analysis is conducted using conjugate prior distribution and likelihood function given empirical hyper-parameters. After derived conditional distributions, Gibbs sampler is then adopted to sample posterior distribution and generate Markov-chain of the parameters and (Section 2.2). The convergence performance of the sampling process is diagnosed (Section 2.3). Based on the statistics evaluation, effectiveness hypothesis is tested for hormone treatment methods (Section 2.4), followed by sensitivity analysis of estimated parameters with different priors in Bayesian method (Section 2.5).

## ***1.3 Major Conclusions***

Based on all the results, three major results/findings are concluded here:

1. Using provided data, the mean estimation of is 1.214, with standard deviation 0.494 and 95% confidence interval [0.536, 2.431].
2. It suggests that hormone treatment doesn’t benefit women who were treated for breast cancer, and it actually do harm to the patient.
3. Parameter is more sensitive to hyper-parameters in priors than , the company should do the simulation using a prior over a large range of hyper-parameters to reduce uncertainty.

# 2. Technic Report

## ***2.1 Basic Statistics of Breast Cancer Data***

All the data collected in the experiments [1] is shown in Table 1. To explore the data, recurrence time and censoring time is summarized for the two groups in terms of mean and standard deviation (See Table 2 and Table 3). Also, the boxplots are depicted (Figure 1 and Figure 2).

|  |  |  |
| --- | --- | --- |
| Table 1. Observed data for breast cancer treatment | | |
| Times | Hormone Treated | Control |
| Recurrence Times | 2 4 6 9 9 9 13 14 18  23 31 32 33 34 43 | 1 4 6 7 13 24 25 35 35 39 |
| Censoring Times | 10 14 14 16 17 18 18 19 20 20  21 21 23 24 29 29 30 30 31 31  31 33 35 37 40 41 42 42 44 46  48 49 51 53 54 54 55 56 | 1 1 3 4 5 8 10 11 13 14 14  15 17 19 20 22 24 24 24 25  26 26 26 28 29 29 32 35 38  39 40 41 44 45 47 47 47 50 50 51 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 2. Basic statistics of recurrence | | |  | Table 3. Basic statistics of censoring | | |
| Recurrence Times | Mean | Standard  Deviation |  | Censoring Times | Mean | Standard  Deviation |
| Hormone Group | 18.67 | 13 |  | Hormone Censoring | 32.79 | 13.74 |
| Control Group | 18.9 | 14.42 |  | Control Censoring | 26.1 | 15.14 |

|  |  |
| --- | --- |
|  |  |
| Figure 1. Boxplots of recurrence | Figure 2. Boxplots of censoring |

As it can be seen, the mean and standard deviation of recurrence for hormone treated group are very close to that of control group. But for censoring time, hormone treated group have greater mean and smaller standard deviation. Since we only have 15 recurrence observation for hormone treated group and 10 for control group, it is reasonable that the data is not symmetric. Indicated by boxplots, the time of hormone treated group is skewed to right.

Furthermore, Kaplan Meier estimator is utilized to estimate survival function on the dataset to describe the patient’s survival probability for a certain time (). The survival function is constructed by computing the cumulative fraction of patients living for a certain amount of time after treated (See Figure 3).

|  |
| --- |
|  |
| Figure 3. Survival function constructed by Kaplan Meier estimator |

It seems that there is no significant different between the two groups in terms of survival probability if considered time period is less than 10 months. And a slightly higher probability is observed for control group when survival time is larger than 10 months.

## ***2.2 Parameter Estimation of Posterior Distribution***

Let be the data for person in the hormone treated group, where is the time and equals 1 if is a recurrence time and 0 if a censored time. The data for the control group can be written similarly and the likelihood function is then given in formula (1).

(1)

It could be obtained from the collected data (Table 1) that , thus, the likelihood function will be:

(2)

Multiply the proposed joint prior distribution (formula (3)) with hyper-parameter , we get the joint posterior distribution in formula (4):

(3)

(4)

Then logit transformation is applied from to by , and the corresponding Jacobian is .

Thus, joint posterior distribution could be rewritten as:

(5)

Then, the sampling target distribution (formula (5)) in log form is obtained as in formula (6):

(6)

Accordingly, the conditional distributions of parameters are derived in formula (7) and (8) which could not be categorized into classic distributions (please note that denotes conditional on all other parameters and data).

(7)

(8)

Up to now, the implementation of mixed Gibbs sampler [2] follows the procedure below:

Step 1: Sample from *U*(0,1), and let ; set ; set all hyper-parameters with the reasonable values suggested by physicians, .

Step 2: Given , sample from ; sample using *M-H* algorithm [3] with a random walk *N*(0, 0.52) proposal distribution because the conditional distribution does not have a closed form. Transfer back to by .

## ***2.3 Convergence Diagnostics of Sampling Process***

To check the effectiveness of above sampling process on the parameters, diagnostics toolkit [4] including sample path, cusum plot and autocorrelation is applied to illustrate the sampling convergence (See Figure 4, Figure 5 and Figure 6).

|  |
| --- |
|  |
| Figure 4. Sample path for , and |
|  |
| Figure 5. Cusum plot for , and |
|  |
| Figure 6. Autocorrelation plot for , and |

For sample path, every 1000 points is plotted and totally 1000,000 iterations are run. First 10% is count as burn-in period. The sample paths of the three parameters indicate it could randomly go up and down through the whole space which suggest mixed well sampling.

In cusum plot, which displays the cumulative sums of the deviations of each sample value from the target value, the curve goes up and down wiggle around 0. In addition, the correlation of each sampling point drops very quickly and stabilize at 0 in autocorrelation plot. Thus it is safe to believe our sample for are all convergent and mixed well.

## ***2.4 Hypothesis Test of Hormone Treatment***

After sampling process for 900,000 (which is sufficiently enough) and convergence checking, the estimation is summarized in Table 4. To better compare prior and estimated posterior distribution for , they are plotted on the same scale in Figure 7.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 4. Parameter estimation results | | | | |
| Parameter | Mean | Standard Deviation | 2.5% quantile | 97.5% quantile |
|  | 0.009 | 0.0027 | 0.005 | 0.015 |
|  | 1.214 | 0.4941 | 0.536 | 2.431 |

|  |
| --- |
|  |
| Figure 7. Prior (curve) and estimated posterior distribution (histogram) for |

As it can be seen in Figure 7, the given prior put more weight on the interval (0,1) especially in (0, 0.5), while most of the posterior are actually in (0.5, 2.4).

Based on the estimation results, hypothesis test [5] is adopted (null hypothesis ) to further decide whether there is significant difference between hormone treated group and control group. Normally distribution could be safely assumed here with a very large sample size (900,000).

Indicated by , null hypothesis is then rejected. Until now, we have sufficient evidence to conclude that there is significant difference between hormone treated group and control group in terms of recurrence time. And since the mean of is 1.24>1, the hormone treatment doesn’t benefit women who were treated for breast cancer, and it actually do harm to breast cancer patient.

## ***2.5 Sensitivity Analysis of Estimation***

Figure 7 has shown some drift between prior and estimated posterior and some common criticism of Bayesian analysis is that the results may highly dependent on the chosen priors. To investigate this issue Gibbs sampler is repeated for half and double the original hyper-parameter values. The results are then recorded in Table 5 and Table 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 5. Parameter estimation results of under different hyper-parameter settings | | | | |
| Parameter | Mean | Standard Deviation | 2.5% quantile | 97.5% quantile |
| Original | 1.214 | 0.4941 | 0.536 | 2.431 |
| Half | 1.311 | 0.5557 | 0.5618 | 2.6940 |
| Double | 1.059 | 0.4017 | 0.4899 | 2.0380 |
| Table 6. Parameter estimation results of under different hyper-parameter settings | | | | |
| Parameter | Mean | Standard Deviation | 2.5% quantile | 97.5% quantile |
| Original | 0.009 | 0.0027 | 0.005 | 0.015 |
| Half | 0.009 | 0.0026 | 0.0043 | 0.0146 |
| Double | 0.010 | 0.0028 | 0.0057 | 0.0165 |

Shown by the results, half or double the hyper-parameter value basically doesn’t affect the estimation of parameter while it could cause as much as 0.3 (24.7% of the original mean) and 0.154 (31.2% of the original standard deviation) on the mean and standard deviation estimation for parameter . Therefore, to reduce the uncertainty, it is recommended that the company could use a prior with a large range of hyper-parameters value and adopted the mean of all the estimated results as possible final results.

# 3. Code in R

## ***3.2 Gibbs Sampler***

library(MCMCpack)

data1=c(2,1,4,1,6,1,9,1,9,1,9,1,13,1,14,1,18,1,23,1,31,1,32,1,33,1,34,1,43,1,10,0,14,0,14,0,16,0,17,0,18,0,18,0,19,0,20,0,20,0,21,0,21,0,23,0,24,0,29,0,29,0,30,0,30,0,31,0,31,0,31,0,33,0,35,0,37,0,40,0,41,0,42,0,42,0,44,0,46,0,48,0,49,0,51,0,53,0,54,0,54,0,55,0,56,0)

yh=matrix(data = data1,53,2,byrow = T) # data of hormone treatment group

data2=c(1,1,4,1,6,1,7,1,13,1,24,1,25,1,35,1,35,1,39,1,1,0,1,0,3,0,4,0,5,0,8,0,10,0,11,0,13,0,14,0,14,0,15,0,17,0,19,0,20,0,22,0,24,0,24,0,24,0,25,0,26,0,26,0,26,0,28,0,29,0,29,0,32,0,35,0,38,0,39,0,40,0,41,0,44,0,45,0,47,0,47,0,47,0,50,0,50,0,51,0)

yc=matrix(data = data2, ncol=2,byrow = T) # data of control group

del.h=sum(yh[,2]);del.h # sum over all delta-h

del.c=sum(yc[,2]);del.c # sum over all delta-c

x.h= sum(yh[,1]);x.h # sum over all xi-h

x.c= sum(yc[,1]);x.c # sum over all xi-c

mu.h.1 = sum(yh[,1] \* yh[,2])/sum(yh[,2]) # mean of h group with label 1

mu.h.0 = sum(yh[,1] \* (yh[,2]-1)/sum((yh[,2]-1))) # mean of h group with label 1

mu.c.1 = sum(yc[,1] \* yc[,2])/sum(yc[,2]) # mean of c group with label 1

mu.c.0 = sum(yc[,1] \* (yc[,2]-1)/sum((yc[,2]-1))) # mean of c group with label 1

################ Gibbs

log.post <- function(d,c,b,a,tau,u) {

(26+a)\*u-(27+a)\*log(1+exp(u))+(15+b)\*log(tau)-(1233+c)\*(exp(u)/(exp(u)+1))-

(1526+d)\*(exp(u)/(exp(u)+1))\*tau

}

niter = 1\*10^6 # number of iterations

theta = rep(0,niter) # define (hyper)parameters

u = rep(0,niter)

tau = rep(0,niter)

a = 1.5

b = 0.5

c = 30

d = 60

j = 0 # counter for extra runs to make theta in (0,1) but no (0,1]

l = 0

# Gibbs Sampler Implementation

set.seed(1) # set random seed

theta[1] = runif(1, min = 0, max = 1) # Initialization of parameters

u[1] = log(theta[1]/(1-theta[1]))

tau[1] = rgamma(1,shape=b+1, rate=(theta[1])\*d)

for (i in 2:niter) { # for-loop for Gibbs sampler

tau[i] <- rgamma(1,shape = 16+b, rate = (exp(u[i-1])/(exp(u[i-1])+1))\*(1526+d))

# update theta in (0,1) by M-H algo

temp2=1

while (temp2==1) {

u[i] <- u[i-1]+rnorm(1,0,0.5)

R2 <- exp(log.post(d,c,b,a,tau[i],u[i])

-log.post(d,c,b,a,tau[i],u[i-1]))

if(R2<1){

if(rbinom(1,1,R2)==0) { u[i]=u[i-1]; l = l+1}

}

temp2=exp(u[i])/(1+exp(u[i])) # transform u back to theta

j = j+1

}

theta[i]=temp2

# update tau by rgamma

}

## ***3.3 Sampling Convergence***

#Plots of the output

burn.in=1:(niter/10)

# sample path plot

par(mfrow=c(1,3)) # sample path plot by parts

plot(theta[seq(1,niter,by=1000)],type="l", ylab = '', xlab ='-th 1000 iterations', main = 'sample path for theta')

abline(v=100, lty = 3)

plot(u[seq(1,niter,by=1000)],type="l", ylab = '', xlab ='-th 1000 iterations', main = 'sample path for u')

abline(v=100, lty = 3)

plot(tau[seq(1,niter,by=1000)],type="l", ylab = '', xlab ='-th 1000 iterations', main = 'sample path for tau')

abline(v=100, lty = 3)

# cusum plot for mean of all parameters

theta.temp <- theta[-burn.in]

theta.ntemp <- length(theta.temp)

theta.mtemp <- mean(theta.temp)

theta.temp.1 <- cumsum(theta.temp-theta.mtemp)

u.temp <- u[-burn.in]

u.ntemp <- length(u.temp)

u.mtemp <- mean(u.temp)

u.temp.1 <- cumsum(u.temp-u.mtemp)

tau.temp <- tau[-burn.in]

tau.ntemp <- length(tau.temp)

tau.mtemp <- mean(tau.temp)

tau.temp.1 <- cumsum(tau.temp-tau.mtemp)

par(mfrow=c(1,3))

plot(theta.temp.1[seq(1,niter,by=1000)], ylab ='',type="l",xlab ='-th 1000 iterations', main = 'cusum plot for theta')

abline(h=0, lty=3)

plot(u.temp.1[seq(1,niter,by=1000)], ylab ='',type="l",xlab ='-th 1000 iterations', main = 'cusum plot for u')

abline(h=0, lty=3)

plot(tau.temp.1[seq(1,niter,by=1000)], ylab ='', type="l",xlab ='-th 1000 iterations',col = rgb(0,0,1,0.8), main = 'cusum plot for tau')

abline(h=0, lty=3)

# autocorrelation plot

par(mfrow=c(1,3))

acf(theta.temp[seq(1,niter-100000,by=1000)], lag.max=40, type="correlation", xlab="N", main = 'autocorrelation of theta');

acf(u.temp[seq(1,niter-100000,by=1000)], lag.max=20, type="correlation", xlab="N",col = rgb(1,0,0,0.8), main = 'autocorrelation of u');

acf(tau.temp[seq(1,niter-100000,by=1000)], lag.max=20, type="correlation", xlab="N",col = rgb(1,0,0,0.8), main = 'autocorrelation of tau');

# 4. References

[1] P. Giudici, G. Givens, and B. Mallick, *Computational Statistics (Breast Cancer Data, page 232)*, Second Edi. Wiley, 2013.

[2] P. Giudici, G. Givens, and B. Mallick, *Computational Statistics (Gibbs Sampling, page 209)*, Second Edi. Wiley, 2013.

[3] P. Giudici, G. Givens, and B. Mallick, *Computational Statistics (Metropolis-Hastings Algorithm, page 202)*, Second Edi. Wiley, 2013.

[4] P. Giudici, G. Givens, and B. Mallick, *Computational Statistics (Sample Graphic Diagnostics, page 219)*, Second Edi. Wiley, 2013.

[5] P. Giudici, G. Givens, and B. Mallick, *Computational Statistics (Hypothesis Testing, page 301)*, Second Edi. 2013.