STA623 - Bayesian Data analysis - Assignment 2

28 October - 1 November 2024

Marc Henrion

Assigment

Please email your typed or scanned solutions before 23:59 on Monday 2 December 2024 to BOTH `mhenrion@mlw.mw` and `biostat-unima@unima.ac.mw`.

Please include `STA623 - Assignment 2` in the subject line. Please include your code, model output and graphs. Please comment any submitted code.

Notation

Please try to use the following notation where possible.

- X, Y, Z - random variables
- x, y, z - measured / observed values
- $\bar{X}$, $\bar{Y}, \bar{Z}$ - sample mean estimators for X, Y, Z
- $\bar{x}$, $\bar{y}, \bar{z}$ - sample mean estimates of X, Y, Z
- $\hat{T}$, $\hat{t}$ - given a statistic T, estimator and estimate of T
- $P(A)$ - probability of an event A occurring
- $f_X(.)$, $f_Y(.), f_Z(.)$ - probability mass / density functions of X, Y, Z
- $p(.)$ - used as a shorthand notation for pmfs / pdfs if the use of this is unambiguous
- $X \sim F$ - X distributed according to distribution function F
- $E[X]$, $E[Y]$, $E[Z]$, $E[T]$ - the expectation of X, Y, Z, T respectively

Table 1: Please use the random seed associated with your name / ID. Solutions using other data than those generated using your seed will not be accepted.

Student	ID	Seed
Hlungumazi Ngwira	MSC-BIO-STAT-03-22	2412
Abdul Hamza	MSC-BIO-STAT-14-23	2304
Francisco Kawonga	MSC-BIO-STAT-15-23	824
Blessings Chirambo	MSC-BIO-STAT-18-23	1092
Funny Osward	MSC-BIO-STAT-22-23	1296
Christopher Phiri	MSC-BIO-STAT-23-23	1025
Brian Mtofu	MSC-BIO-STAT-24-23	1344
Tereza Mwanavava	MSC-BIO-STAT-J-01-24	1408
Weldon Chihana	MSC-BIO-STAT-J-03-24	1050
Wongani Luhanga	MSC-BIO-STAT-J-04-24	2321
Joseph Kenneth	MSC-BIO-STAT-J-05-24	1792
Harry Milal	MSC-BIO-STAT-J-08-24	1206
Eneles Mponda	MSC-BIO-STAT-J-10-24	1736
Harriet Mchira	MSC-BIO-STAT-J-11-24	1791
Germue Gbawoquiya	MSC-BIO-STAT-J-17-24	2616
Marion Maganga	MSC-MAT-03-23	2460

Exercise

For the exercise below, you will need to specify a seed value. You will be given individual seed numbers according to the table on the previous page. **You have to use your own individual seed value** – your data (and hence your results) will be unique to you and different from those of your colleagues.

Use the code below (downloadable as file `hospitalWaitTimes_generateData_2024.R` from GitHub) to simulate data on A&E waiting times for several hospitals.

```
set.seed(0000) # REPLACE 0000 with your individual seed value!
# Solutions using the seed value 0000 will not be accepted.

# Generate data
n<-rpois(n=1,lambda=250)
hospRf<-rnorm(n=8,mean=0,sd=0.3)
hospRf<-hospRf-mean(hospRf)

dat<-data.frame(
  PID=paste(sep="","P",24000+1:n),
  sex=sample(c("M","F"),size=n,replace=TRUE,prob=c(0.5,0.5)),
  triage=factor(
    levels=c("Emergency","Priority","Queue"),
    sample(x=c("Emergency","Priority","Queue"),
           size=n,replace=TRUE,prob=c(0.1,0.25,0.65))
  ),
  hospital=factor(
    levels=paste(sep="","H",1:8),
    sample(x=paste(sep="","H",1:8),
           size=n,replace=T,prob=c(0.25,0.15,0.15,rep(0.09,5)))
  )
) %>%
  dplyr::mutate(
    hospRanEf=hospRf[as.integer(hospital)],
    wait=rexp(n=n,
              rate=0.75
              +case_when(triage=="Emergency"~rnorm(n=1,mean=1.5,sd=0.25),
                         triage=="Priority"~rnorm(n=1,mean=0.25,sd=0.05),
                         triage=="Queue"~0)
              +hospRanEf)
  ) %>%
  dplyr::select(!hospRanEf)
```

The dataset you just simulated contains the following columns:

- `pid` - this is just an anonymised patient identification number
- `sex` - this records the biological sex of each patient
- `triage` - records the category that the patients were triaged into by an admission nurse (`emergency`, `priority` or `queue`); the idea is that emergencies get seen without delay, priority cases get seen more quickly than normal cases and then the third category is for all other cases
- `hospital` - this records an identification code for the hospital where each patient was seen
- `wait` - this records the waiting time (in hours) that each patient had to wait before being seen by a A&E doctor

Use JAGS to fit the following model, choosing priors of your own choosing for each parameter, writing $Y_{i,j}$ for the waiting time variable for patient $i = 1, \ldots, n$ seen in hospital $j = 1, \ldots, k$:

$$Y_{i,j} = \beta_0 + \beta_1 \cdot male\_sex_i + \beta_2 \cdot triage\_emergency_i + \beta_3 \cdot triage\_priority + \mu_j + \epsilon_i$$

where $\mu_j \sim \mathcal{N}(0, \rho^2), j = 1, \ldots, k$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$

1. List the number of observations and the average waiting time for your particular dataset. [5 marks]
2. Explain the choice of prior distributions for all model parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \rho^2, \sigma^2)$. [15 marks]
3. Write JAGS model code to fit the model. [35 marks]
4. Fit the model, then show and summarise (as a point estimate + confidence interval) the posterior distributions for the various parameters. Explain your choice of Bayesian estimators you report. [15 marks]
5. Show trace plots and histograms for all model parameters and compute the effective sample size and Gelman-Rubin potential scale reduction factors. Discuss the results you are getting. [20 marks]
6. Discuss other model checks you could do. [10 marks]