Module 6: Bayesian Methods for Clinical Research - Computational Methods & Application

Rebecca Hubbard, Lurdes Inoue

July 25, 2017

Install R

- Go to http://cran.rstudio.com/ (http://cran.rstudio.com/)
- Click on the "Download R for [operating system]" link that is appropriate for your operating system and follow the instructions.
- Open R and make sure it works (i.e. that no error messages come up)

Install RStudio

- Go to http://www.rstudio.com/products/rstudio/download/ (http://www.rstudio.com/products/rstudio/download/)
- Select the installer that is appropriate for your operating system under "Installers for Supported Platforms" and follow the instructions.
- Open RStudio and make sure it works.

Install R Packages

- For this module we will be using the *INLA*, *rjags*, and *coda* packages for Bayesian estimation and MCMC convergence diagnostics
- We will also be using the *eha* and *survival* packages for classical survival regression
- To use these packages you first need to install them using *install.packages()*

```
install.packages("INLA", repos="https://www.math.ntnu.no/inla/R/stable")
install.packages("rjags")
install.packages("coda")
install.packages("eha")
install.packages("survival")
```

Install JAGS

- You will also need to install JAGS

- Go to https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/ (https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/)
- Select the version that is appropriate for your operating system, and download and install the software

Load libraries

- From within R Studio you will need to load the following libraries:

```
library(INLA)
library(rjags)
library(coda)
library(eha)
library(survival)
```

- After the first time you install the packages on your computer, you will only need to load the libraries in the future

---

Bayesian GLMs using INLA

In this lab, we will conduct an analysis using Bayesian logistic and survival regression models estimated with INLA. We will use data from the Western Collaborative Group Study (WCGS), a study of the association between cardiovascular health and behavioral pattern conducted in a cohort of male volunteers. Subjects were recruited in 1960-1961 and followed for up to 9 years for onset of coronary heart disease (CHD). The scientific question of interest is whether behavioral pattern is associated with CHD. Specifically, investigators hypothesized that men with a "Type A" behavioral pattern would be more likely to experience CHD. Because this is an observational study design it is important to account for confounding due to many factors such as cigarette smoking and elevated BMI.

We will use the following variables from this data set:

- age Age: age in years
- behpat Behavior pattern: (A1, A2, B3, B4)
- bmi Body mass index
- chd Indicator of CHD at any time during follow-up: 0 = no; 1 = yes
- chd01 Coronary heart disease within 5 years: 0 = no; 1 = yes
- chol Cholesterol: mg/100 ml
- dbp Diastolic blood pressure: mm Hg
- dibpat Dichotomous behavior pattern: 0 = Type B; 1 = Type A
- height Height: height in inches

- id Subject ID
- ncigs Smoking: Cigarettes/day
- sbp Systolic blood pressure: mm Hg
- time Time in days from baseline to onset of CHD
- smoke: No = non-smoker; Yes = current smoker
- weight Weight: pounds

You can download the data file and read it into R as follows:

```
wcgs <- read.csv("https://raw.githubusercontent.com/rhubb/SISCR2017/master/data/wcgs.csv", header = T)
```

1. We will start by using logistic regression to analyze the association between behavioral pattern and onset of CHD within 5 years of baseline (*chd01*). Begin by conducting an exploratory data analysis to summarize the distribution of CHD, behavioral pattern, and possible confounders included in the data set. What conclusions do you reach regarding the role that number of cigarettes, BMI, and age may play in the analysis of the association between CHD and behavioral pattern?

```
#-- univariate tables for categorical variables
table(wcgs$chd01)/sum(table(wcgs$chd01))
table(wcgs$behpat)/sum(table(wcgs$behpat))
table(wcgs$smoke)/sum(table(wcgs$smoke))

#-- bivariate tables for categorical predictors and CHD
# CHD and behavioral pattern
table(wcgs$behpat,wcgs$chd01)
t(sweep(table(wcgs$chd01,wcgs$behpat),2,rowSums(table(wcgs$behpat,wcgs$chd01)),"/"))
# CHD and smoking
table(wcgs$smoke,wcgs$chd01)
t(sweep(table(wcgs$chd01,wcgs$smoke),2,rowSums(table(wcgs$smoke,wcgs$chd01)),"/"))
# Behavioral pattern and smoking
table(wcgs$behpat,wcgs$smoke)
t(sweep(table(wcgs$smoke,wcgs$behpat),2,rowSums(table(wcgs$behpat,wcgs$smoke)),"/"))

#-- summary statistics for continuous variables by CHD
# Age
tapply(wcgs$age,wcgs$chd01,mean)
tapply(wcgs$age,wcgs$chd01,sd)
boxplot(wcgs$age ~ wcgs$chd01, xlab = "CHD", ylab = "Age (years)")

# BMI
tapply(wcgs$bmi,wcgs$chd01,mean)
tapply(wcgs$bmi,wcgs$chd01,sd)
boxplot(wcgs$bmi ~ wcgs$chd01, xlab = "CHD", ylab = "BMI")

# Number of cigarettes
tapply(wcgs$ncigs,wcgs$chd01,mean)
tapply(wcgs$ncigs,wcgs$chd01,sd)
boxplot(wcgs$ncigs ~ wcgs$chd01, xlab = "CHD", ylab = "Number of Cigarettes")
```

2. Next, use Bayesian logistic regression to analyze the association between CHD and behavioral pattern, accounting for possible confounders based on your results from (1). We will explore results using several prior distributions. For each prior distribution can you think of a context in which this prior would be preferred? Compare your results to a classical logistic regression. How does the interpretation of the results differ for the Bayesian GLM compared to the frequentist GLM?

a. Normal(0,10) priors

```
# -- Normal priors for regression coefficients (with mean=0 and scale=10)
chd.n10 <- inla(chd01~ factor(behpat) + smoke + age, data=wcgs, family = "binomial",
     control.fixed=list(mean.intercept=c(0),prec.intercept=c(1/10),mean=c(0,0),prec=rep
(1/10,2)))

chd.n10$summary.fix
# -- Plot posterior densities
plot(chd.n10, plot.prior = TRUE)
```

b. Normal(0,0.1) priors

```
# -- Normal priors for regression coefficients (with mean=0 and scale=0.1), N(0,10) pri
or for intercept
chd.n01 <- inla(chd01~ factor(behpat) + smoke + age, data=wcgs, family = "binomial",
     control.fixed=list(mean.intercept=c(0),prec.intercept=c(1/10),mean=rep(0,5),prec=r
ep(10,5)))

chd.n01$summary.fix

# -- Plot posterior densities
plot(chd.n01, plot.prior = TRUE)
```

c. Classical logistic regression

```
chd.glm1 <- glm(chd01~ factor(behpat) + smoke + age, data=wcgs, family=binomial)
summary(chd.glm1)
```

3. Using one of the Bayesian models you fit in (2), interpret your results. What do you conclude about the association between behavioral pattern and CHD? Does your choice of prior affect your conclusions? How does the interpretation of results for the Bayesian GLM differ from the results of the frequentist GLM?

```
# -- Exponentiate results to obtain odds ratios
exp(chd.n10$summary.fix)
```

4. Since some individuals were censored prior to the end of follow-up, a more appropriate way to analyze these data is with survival analysis. Analyze the association between time to onset of CHD and behavioral pattern, adjusting for the same confounders used in your logistic regression model using:

a. Cox proportional hazards model

```
chd.cph <- coxph(Surv(time, chd) ~ factor(behpat) + smoke + age, data=wcgs)
summary(chd.cph)
```

b. Parametric survival regression

```
chd.weib <- phreg(Surv(time, chd) ~ factor(behpat) + smoke + age, data=wcgs, dist = "weibull")
summary(chd.weib)
```

c. Bayesian non-parametric survival model

```
chd.np <- inla(inla.surv(time, chd) ~ factor(behpat) + smoke + age, family="coxph",data=wcgs,control.hazard=list(model="rw1", n.intervals=10))
summary(chd.np)
exp(chd.np$summary.fix)

# plot baseline hazard function
plot(chd.np$summary.random$baseline.hazard[,"ID"],
 exp(chd.np$summary.fixed[1,1]+chd.np$summary.random$baseline.hazard[,"mean"]), type="S",
 xlab = "time", ylab = "Baseline hazard")
```

d. Bayesian parametric survival model

```
chd.weib <- inla(inla.surv(time, chd) ~ factor(behpat) + smoke + age, family="weibullsurv",data=wcgs)
summary(chd.weib)
exp(chd.weib$summary.fix)
```

e. What do you conclude about the relationship between behavioral pattern and hazard of CHD? Are your conclusions affected by which estimation approach you chose? If so, which one would you prefer in this context and why?

Analysis of Correlated Data

In this lab, we will conduct an analysis using Bayesian hierarchical models estimated with *INLA* and *rjags*. We will use data from the Study of Osteoporotic Fractures, a longitudinal study of fractures and falls in older women in the US. The study investigated risk factors associated with fractures and falls as well as changes over time in bone mineral density (BMD), an early indicator of changes in bone strength that may precede osteoporotic fractures. We will use data from this study to investigate associations between BMD, body mass index (BMI), instrumental activities of daily living (IADL), and age at menopause.

We will use the following variables from this data set:

- id Patient id
- visit Visit number (continuous)
- totbmd Bone mineral density (continuous)
- bmi Body mass index (continuous)
- n_iadl Numer of impaired instrumental activities of daily living
- age_base Age at baseline (continuous)
- meno_age Age at menopause (continuous)
- dbp Diastolic blood pressure: mm Hg
- base_totbmd Bone mineral density at baseline (continuous)

You can download the data file and read it into R as follows:

```
sof <- read.csv("https://raw.githubusercontent.com/rhubb/SISCR2017/master/data/sof3.csv", header = T)
```

1. Conduct an exploratory analysis of longitudinal changes in BMD using descriptive statistics and plots. How many observations are available for each woman? On average how much does BMD change over time?

```
# Number of women in the data set
length(unique(sof$id))

# Distribution of number of non-missing BMD measures available per woman
summary(c(table(sof$id[!is.na(sof$totbmd)])))
barplot(table(table(sof$id[!is.na(sof$totbmd)])), xlab = "Number BMD measures", ylab =
"Number of women")

# Correlation between visit number and BMD
cor(sof$visit,sof$totbmd, use = "pairwise.complete.obs")

# Plot of BMD across visits for first 100 women
ptid <- unique(sof$id)
plot(sof$visit[sof$id==ptid[1]],sof$totbmd[sof$id==ptid[1]], xlab = "Visit Number",
 ylab = "BMD", ylim = c(0.1,1.8), type = "l", col = "grey")
for (i in 2:100){
lines(sof$visit[sof$id==ptid[i]],sof$totbmd[sof$id==ptid[i]], col = "grey")
}

# Simple linear regression analysis of change in BMD over time
summary(lm(totbmd ~ visit, data = sof))
```

2. Since repeated BMD measurements made for the same woman are likely to be highly correlated, a formal analysis of change in BMD over time needs to account for within-woman correlation. This can be achieved using a Bayesian hierarchical regression model. In this model we will assume that multiple measurements made for the same woman are exchangeable conditional on subject-specific mean parameter θ_i and that these subject-specific means arise from a common distribution with hyperparameter μ. Consider adding additional predictors to this model that may help to explain variation in BMD. What do you conclude about longitudinal trends in BMD?

```
mod1 <- inla(totbmd ~ visit + f(id, model = "iid"),family="gaussian", data = sof)
summary(mod1)

mod2 <- inla(totbmd ~ visit + bmi + age_base + f(id, model = "iid"),family="gaussian",
data = sof)
summary(mod2)
```

3. Next we will repeat this analysis using rjags to implement an MCMC estimation method.

a. Prepare data for use by JAGS. Be sure to include any variables you would like to incorporate into your regression model.

```
# first we need to remove observations with missing data
sof.nomiss <- na.omit(sof[,c("id","totbmd","visit")])

# next create a list containing the data elements that will be used by JAGS
sof.jags <- list(totbmd = sof.nomiss$totbmd, visit = sof.nomiss$visit, id = rep(seq(1,length(unique(sof.nomiss$id))),times=table(sof.nomiss$id)), n = length(sof.nomiss$id), m = length(unique(sof.nomiss$id)))
```

b. Write a JAGS model specifying the likelihood and priors.

```
sof.model <- "
model{
  ## likelihood
  for (i in 1:n){
    totbmd[i] ~ dnorm(mu[i],tausq)
    mu[i] <- beta0 + beta1*visit[i] + b0[id[i]]
  }

  ## priors
  for (j in 1:m){
    b0[j] ~ dnorm(0,taubsq)
  }
  beta0 ~dnorm(0,0.01)
  beta1 ~dnorm(0,0.01)
  tausq <- 1/sigmasq
  taubsq <- 1/sigmabsq
  sigmasq  ~ dunif(0,1)
  sigmabsq ~ dunif(0,1)
}
"
```

c. Use JAGS to estimate regression parameters and subject-specific random effects

```
# define jags model
mod <- jags.model(file=textConnection(sof.model), data=sof.jags, inits=list(beta0 = 0,
beta1 = 0, b0 = rep(0,length(unique(sof.nomiss$id))), sigmasq = 0.01, sigmabsq = 0.01),
n.chains=2, n.adapt=10000)

# specify parameters to be monitored
params <- c("beta0","beta1","sigmasq","sigmabsq")

# run jags and save posterior samples
samps <- coda.samples(mod, params, n.iter=10000, n.thin = 10)

# summarize posterior samples
summary(samps)
```

c. Use diagnostic plots and statistics to evaluate model convergence. Are there any parameters for which it appears that your chains have not converged? Can you think of any approaches that might improve the convergence of your chains?

```
# Traceplots
traceplot(samps)

# Auto-correlation plots
autocorr.plot(samps)

# Gelman & Rubin diagnostics
gelman.diag(samps)
gelman.plot(samps)

# Geweke diagnostic
geweke.diag(samps)

# Raftery & Lewis diagnostic
raftery.diag(samps)

# Heidelberger & Welch diagnostic
heidel.diag(samps)
```

4. How do your results obtained using *JAGS* compare to those from *INLA*? Using, JAGS try out several different prior distributions. Are results sensitive to your choice of prior?

Meta-Analysis

In this lab, we will conduct a meta-analysis of 28 studies investigating the effect of interventions designed to reduce cholesterol on ischemic heart disease (IHD). The outcome of interest in these studies (IHD) was occurrence of fatal or non-fatal myocardial infarction.

We will use the following variables from this data set:

- id Trial id
- cholreduc Average cholesterol reduction in treated group - average reduction in control grup (mmol/l)
- Y Number of IHD events
- N Total number of participants
- Trt Treatment group: 1 = Intervention, 0 = Control

You can download the data file and read it into R as follows:

```
chol <- read.csv("https://raw.githubusercontent.com/rhubb/SISCR2017/master/data/cholesterol.csv", header = T)
```

1. We will conduct a series of Bayesian meta-analyses of these data to investigate the association between interventions targeting cholesterol reduction and odds of IHD.

a. Meta-analysis ignoring between-trial heterogeneity

```
ma1 = inla(Y~ factor(Trt), data=chol, Ntrials=N, family="binomial")
summary(ma1)
# posterior median odds ratio
exp(ma1$summary.fixed[2,4])
```

b. Meta-analysis ignoring between-trial heterogeneity accounting for reductions in cholesterol level

```
# created centered cholesterol reduction variable
chol$chol.cent <- chol$ cholreduc-mean(chol$ cholreduc)

ma2 = inla(Y~ factor(Trt) + chol.cent, data=chol, Ntrials=N, family="binomial")
summary(ma2)
# posterior median odds ratio
exp(ma2$summary.fixed[2,4])
```

c. Meta-analysis accounting for between trial heterogeneity via fixed-effects

```
ma3 = inla(Y~ -1 + factor(id)+ factor(Trt), data=chol, Ntrials=N, family="binomial")
summary(ma3)
# posterior median odds ratio
exp(ma3$summary.fixed[29,4])
```

d. Meta-analysis accounting for between trial heterogeneity via random-effects

```
ma4 = inla(Y~ factor(Trt) + f(id, model = "iid", param = c(0.001,0.001)), data=chol, Ntrials=N, family="binomial")
summary(ma4)
# posterior median odds ratio
exp(ma4$summary.fixed[2,4])
```

e. Meta-analysis accounting for between trial heterogeneity via random-effects and cholesterol reduction

```
ma5 = inla(Y~ factor(Trt) + chol.cent + f(id, model = "iid", param = c(0.001,0.001)), data=chol, Ntrials=N, family="binomial")
summary(ma5)
# posterior median odds ratio
exp(ma5$summary.fixed[2,4])
```

f. Meta-analysis accounting for between trial heterogeneity via random-effects and allowing for moderation by cholesterol reduction

```
ma6 = inla(Y~ factor(Trt)*chol.cent + f(id, model = "iid", param = c(0.001,0.001)), data=chol, Ntrials=N, family="binomial")
summary(ma6)
# posterior median odds ratio at mean cholesterol reduction
exp(ma6$summary.fixed[2,4])
# effect of 1 mmol/l greater reduction in cholesterol on posterior median odds ratio
exp(ma6$summary.fixed[4,4])
```

2. Overall what do you conclude about the association between cholesterol reduction and IHD? Which meta-analysis approach do you prefer in this example and why?