Module 2: Bayesian Methods for Clinical Research - Introduction

Rebecca Hubbard, Lurdes Inoue

July 24, 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 *LearnBayes* and *arm* packages
- To use these packages you first need to install them using *install.packages()*

```
install.packages("LearnBayes")
install.packages("arm")
```

- You then need to load these libraries:

```
library(LearnBayes)
library(arm)
```

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

Introduction to Bayesian Computing

1. You are conducting a trial of a new treatment for thyroid cancer and have obtained results for the first 80

patients.63 patients responded to treatment and 17 patients did not. We now want to obtain the posterior distribution for the probability of responding to this new treatment. Use R to generate the posterior distribution and compute the posterior median under the following priors:

a. Flat prior on [0,1]

```
#-- prior values for probability of response
theta <- seq(0,1,length.out=99)

#-- likelihood times prior
prior <- rep(1/99, 99)
product <- dbinom(x=63, size=80, prob=theta)*prior

#-- posterior is the normalized likelihood times prior
posterior <- product/sum(product)

#-- plot posterior distribution
plot(theta, posterior, type='h', xlab=expression(~theta))

#-- posterior mean
mean.post <- sum(theta*posterior)

#-- cumulative posterior distribution
cumulative.post <- cumsum(posterior)

#-- median (approximate)
median.post <- theta[max(which(cumulative.post <=0.50))]
```

b. Flat prior on [0.8,1]

```
#-- prior values for probability of response
theta <- seq(0.8,1,length.out=99)

#-- likelihood times prior
prior <- rep(1/99, 99)
product <- dbinom(x=63, size=80, prob=theta)*prior

#-- posterior is the normalized likelihood times prior
posterior <- product/sum(product)

#-- plot posterior distribution
plot(theta, posterior, type='h', xlab=expression(~theta))

#-- posterior mean
mean.post <- sum(theta*posterior)

#-- cumulative posterior distribution
cumulative.post <- cumsum(posterior)

#-- median (approximate)
median.post <- theta[max(which(cumulative.post <=0.50))]
```

   c. What is the danger of using the prior in (b)?

2. Let's analyze the same set of results using *LearnBayes*.

   a. Flat prior on [0,1]

```
triplot(prior=c(1,1),data=c(63,17), where = "topleft")
```

   b. Beta(4,4) prior

```
triplot(prior=c(4,4),data=c(63,17), where = "topleft")
```

3. Now suppose we want to test the hypothesis that the probability of response to our new treatment is 0.8. Assuming we have equipoise regarding this hypothesis (i.e., we think it is equally likely to be true or false) and that our alternative hypothesis is that the response probability is equally likely to take any value from 0 to 1, what are the posterior probability of the null hypothesis and the Bayes Factor? Is there evidence for or against our null hypothesis?

```
pbetat(p0=0.8, prob=0.5, ab=c(1,1), data=c(63,17))
```

4. Finally, what if we want to estimate the posterior predictive distribution of response for a new patient based on the data that we have observed and assuming that we initially had a Beta(4,4) prior for the probability of response. (Recall that by using a conjugate prior our posterior distribution is also Beta with parameters $a + y - 1$ and $b + n - y - 1$.)

    ```
    pbetap(ab=c((4+63-1),(4+17-1)), n=1, s=0:1)
    ```

Bayesian Regression Models

In this lab, we will conduct an analysis using a Bayesian logistic regression model. Data come from a cross-sectional study of 1,225 smokers over the age of 40. Each participant was assessed for chronic obstructive pulmonary disease (COPD), and characteristics of the type of cigarette they most frequently smoke were recorded. The objective of the study was to identify associations between COPD diagnosis and cigarette characteristics.

We will use the following variables from this data set:

- TYPE: Type of cigarette, 1 = Menthol, 0 = Regular
- NIC: Nicotine content, in mg
- TAR: Tar content, in mg
- LEN: Length of cigarette, in mm
- FLTR: 1 = Filter, 0 = No filter
- copd: 1 = COPD diagnosis, 0 = no COPD diagnosis

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

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

1. First carry out some exploratory data analysis to summarize the distribution of copd, cigarette type, nicotine content, and filter.

```
#-- univariate tables for categorical variables
table(copd$copd)/sum(table(copd$copd))
table(copd$TYPE)/sum(table(copd$TYPE))
table(copd$FLTR)/sum(table(copd$FLTR))

#-- bivariate tables for categorical predictors and copd
table(copd$TYPE,copd$copd)
t(sweep(table(copd$copd,copd$TYPE),2,rowSums(table(copd$TYPE,copd$copd)),"/"))
table(copd$FLTR,copd$copd)
t(sweep(table(copd$copd,copd$FLTR),2,rowSums(table(copd$FLTR,copd$copd)),"/"))

#-- summary statistics for nicotine content by copd
by(copd$NIC,copd$copd,mean)
by(copd$NIC,copd$copd,sd)
boxplot(copd$NIC ~ copd$copd, xlab = "COPD", ylab = "Nicotine (mg)")
```

2. Next, use Bayesian logistic regression to analyze the association between COPD and predictors cigarette type, nicotine content, and filter. 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 standard frequentist logistic regression. How does the interpretation of the results differ for the Bayesian GLM compared to the frequentist GLM?

a. Normal(0,10) prior

```
# -- Normal priors for regression coefficients (with mean=0 and scale=10)
copd.n10 <- bayesglm(copd ~ TYPE + NIC + FLTR, data=copd, family=binomial, prior.mean=0
,
   prior.scale=10, prior.df=Inf)
display(copd.n10)
```

b. Normal(0,0.1) prior

```
# -- Normal priors for regression coefficients (with mean=0 and scale=0.1)
copd.n1 <- bayesglm(copd ~ TYPE + NIC + FLTR, data=copd, family=binomial, prior.mean=0,
   prior.scale=0.1, prior.df=Inf)
display(copd.n1)
```

c. Cauchy prior

```
# -- Cauchy priors for regression coefficients
copd.cau <- bayesglm(copd ~ TYPE + NIC + FLTR, data=copd, family=binomial, prior.mean=0,
   prior.scale=10)
display(copd.cau)
```

d. Frequentist logistic regression

```
copd.glm1 <- glm(copd ~ TYPE + NIC + FLTR, data=copd, family=binomial)
summary(copd.glm1)
```

3. Using one of the Bayesian models you fit in (2), interpret your results. How does the interpretation of the Bayesian logistic regression results differ from the classical results in (d)?

4. How would the results in (2) differ if our data set had been smaller? Refit the models in (2) using only the first 200 observations in the data set. What can you say about the results for cigarette type?

a. Normal(0,10) prior

```
# -- Normal priors for regression coefficients (with mean=0 and scale=10)
copd.n10.v2 <- bayesglm(copd ~ TYPE + NIC + FLTR, data=copd[1:200,], family=binomial, prior.mean=0,
   prior.scale=10, prior.df=Inf)
display(copd.n10.v2)
```

b. Normal(0,0.1) prior

```
# -- Normal priors for regression coefficients (with mean=0 and scale=0.1)
copd.n1.v2 <- bayesglm(copd ~ TYPE + NIC + FLTR, data=copd[1:200,], family=binomial, prior.mean=0,
   prior.scale=0.1, prior.df=Inf)
display(copd.n1.v2)
```

c. Cauchy prior

```
# -- Cauchy priors for regression coefficients
copd.cau.v2 <- bayesglm(copd ~ TYPE + NIC + FLTR, data=copd[1:200,], family=binomial, prior.mean=0,
   prior.scale=10)
display(copd.cau.v2)
```

d. Frequentist logistic regression

```
copd.glm1.v2 <- glm(copd ~ TYPE + NIC + FLTR, data=copd[1:200,], family=binomial)
display(copd.glm1.v2)
```

Interim Monitoring of Clinical Trials

In this lab, we will examine data from a trial and determine whether or not to stop the trial at a series of interim analyses based on the posterior probability of success. We will use the function *PP()*, provided below, to simulate the posterior probability of success:

```
# n.total = total sample size for the trial
# nullp = value for response probability under the null hypothesis
# eta = posterior probability of p>nullp must exceed eta to stop the trial
# data = number of successes and failures observed
# prior.par = parameters of Beta prior for p
# B = number of samples from posterior distribution

PP <- function(n.total, nullp, eta=0.95, data=c(12,8), prior.par=c(1,1), B=1000){
  # posterior
  post.par <- data + prior.par

  # samples from posterior distribution
  post.sample <- rbeta(B, post.par[1], post.par[2])

  # samples new values of x (extending to the maximum sample size)
  x.new <- rbinom(B, size=n.total-sum(data), post.sample)

  # organize data with first column number of 'responses' and second 'non responses'
  data.new <- cbind(x.new, n.total-sum(data)-x.new)

  # posterior parameters given predicted data
  post.pred.par <- cbind(data.new[,1] + post.par[1], data.new[,2]+ post.par[2])

  # posterior probability that P(p > nullp |data)
  post.pred <- pbeta(nullp, post.pred.par[,1], post.pred.par[,2], lower.tail=FALSE)

  # posterior predictive probability of success
  PP <- mean(post.pred > eta)
  return(PP)
}
```

In all interim analyses, assume that there is an existing treatment with response probability of 0.3 and that we will declare the trial to be a success if the posterior probability that the therapy under study has a higher response probability than the existing therapy exceeds 0.9. The table below provides the number of success and failures at three interim time points and at the conclusion of the study

1. Using a flat prior (Beta(1,1)), would you decide to stop the trial at any of the interim analyses? Are your interim analysis decisions consistent with the conclusion you would draw at the end of the trial?

```r
#-- interim analyses
PP(n.total = 100, nullp = 0.3, eta = 0.9, data = c(17,22),prior.par = c(1,1))
PP(n.total = 100, nullp = 0.3, eta = 0.9, data = c(21,39),prior.par = c(1,1))
PP(n.total = 100, nullp = 0.3, eta = 0.9, data = c(30,50),prior.par = c(1,1))

#-- posterior probability of response probability exceeding 0.3 at end of trial
pbeta(0.3,(34-1),(66-1), lower.tail = FALSE)
```

2. Repeat (1) using a prior centered at the null hypothesis such as a Beta(0.3,1). How does this affect your conclusions about whether or not to stop the trial?

```r
#-- interim analyses
PP(n.total = 100, nullp = 0.3, eta = 0.9, data = c(17,22),prior.par = c(0.3,1))
PP(n.total = 100, nullp = 0.3, eta = 0.9, data = c(21,39),prior.par = c(0.3,1))
PP(n.total = 100, nullp = 0.3, eta = 0.9, data = c(30,50),prior.par = c(0.3,1))

#-- posterior probability of response probability exceeding 0.3
pbeta(0.3,(34-1),(66-1), lower.tail = FALSE)
```

3. If you were designing a trial would you favor the flat prior, a prior centered at the response probability for the existing therapy, or a different prior? What considerations contribute to your decision?