PB HLTH 250C: Assignment 3

Due Thursday 14 March 2024 at 11:59pm via Gradescope

Read all questions carefully before answering. You may work in small groups of no more than 3 individuals and turn in a single assignment (and everyone in the group will receive the same grade). Work through the entire assignment individually first, then come together to discuss and collaborate. Please maintain numbering on sub-questions, type your responses, and **please keep answers brief.**

Bayesian logistic regression model

We wish to estimate a logistic regression model in a Bayesian framework to assess the relationship between the outcome of cardiovascular disease over 24 years (`cvd`) and smoking, controlling for age, sex, and education.

For a Bayesian analysis, we specify the sampling distribution (the regression model for the outcome), and the prior distribution (the distribution that encodes our beliefs about the unknown parameters):

1. **Sampling distribution:** a logistic regression model for binary CVD status (outcome) as a function of covariates:

$$\begin{aligned} \text{logit}(\pi_i) &= \beta_1 + cursmoke_i\beta_2 + age_i\beta_3 + male_i\beta_4 + educ2_i\beta_5 + educ3_i\beta_6 + educ4_i\beta_7 \\ cvd_i &\sim \text{Binomial}(\pi_i, 1) \end{aligned} \tag{1}$$

   where π_i is the probability of having incident CVD over the follow-up, `cursmoke` is current smoking status (smoker vs. non-smoker), age is in years, `male` is an indicator for male sex (female is reference), `educ2-educ4` are indicators for education categories (high school graduate, some college, college graduate, vs < high school graduate (ref)). The parameters β are the log-odds ratios for the corresponding coefficient.[1]

2. **Prior distribution:** we specify a Normal (Gaussian) prior for the regression parameters (β):

$$\beta_j \sim \text{N}(\mu_j, \tau_j) \tag{2}$$

   where μ_j is the prior mean, and τ_j is the prior precision (1/variance) for coefficient j.

We will specify this in JAGS and conduct an analysis exploring different parameterizations on the prior for smoking.

Tasks:

Load required packages and read data:

```r
library(R2jags)
library(coda)
require(foreign)

load("CVD_data.Rdata")
```

[1] My description here is more concise than I usually require, for brevity's sake.

- Completing the code below, characterize the posterior distribution of the parameters of a logistic regression model for having CVD as a function of current smoking status, age, sex, and education. Our prior assumption is that the slope parameters in the regression model are independent and normally distributed. For now, we assume a vague prior with zero mean and variance 1000 (precision=$\tau = 1000^{-1}$).

Do a bit of variable recoding:

We center (de-mean) and scale (divide by standard deviation) the age variable (as it is continuous), which can help convergence.

```
# Extract data elements from data frame
cvd <- CVD.data$cvd # Outcome
cursmoke <- CVD.data$cursmoke # Exposure (smoking status)
age.c <- as.numeric(scale(CVD.data$age))
male <- as.integer(CVD.data$sex=="male")

# Create education indicators (a shortcut using the model.matrix command)
X.educ <- model.matrix(~-1 + factor(educ), data=CVD.data)
educ1 <- X.educ[,1] # Unused in our analysis (reference category)
educ2 <- X.educ[,2]
educ3 <- X.educ[,3]
educ4 <- X.educ[,4]
```

Create the JAGS code that defines the posterior distribution:

Complete the following function to define the posterior of the model parameters:

```
# JAGS code for the posterior distribution:
model.posterior <- function() {
  for (i in 1:N) {

      # COMPLETE THE EXPRESSION FOR THE PROBABILITY OF THE OUTCOME
      # AS A FUNCTION OF INPUT VARIABLES AND MODEL PARAMETERS

      cvd[i] ~ dbin(pi[i], 1); # Distribution of outcome
    }

  # PRIORS ON BETAS
  for (j in 1:Nx){
      b[j] ~ dnorm(mu[j], tau[j]);  # Independent normal priors
      OR[j] <- exp(b[j]);           # Calculate the odds ratios
  }
}
```

Define elements for the JAGS function:

NOTE: complete specification of mu and tau own own:

```
# Constants to be passed in
N <- length(cvd);           # Number of observations to loop over
Nx <- 7;                    # Number of parameters (w/ intercept)
```

```
n.iter <- 10000;            # Number of iterations to run (total)

# Parameters on the priors:

# COMPLETE SPECIFICATION OF MU AND TAU ON OWN

# List of data elements to pass in
data.list <- list("N", "Nx", # Model constants (# obs, # vars)
                  "cvd", "age.c", # Variable names (next 3 lines)
                  "male", "cursmoke",
                  "educ2", "educ3", "educ4",
                  "mu","tau") # Hyperparameters

# List of parameters to keep track of:
parameters.model <- c("b", "OR")

# Function to randomly generate initial values for each chain:
inits.model <- function() {list (b=rnorm(Nx, 0, sd=.5))}
```

Run the MCMC Algorithm and summarize

```
set.seed(123)
jags.samples <- jags(data=data.list,
                       model.file=model.posterior,
                       inits=inits.model,
                       parameters.to.save=parameters.model,
                       n.iter=n.iter, n.chains=3)
print(jags.samples,digits=4)
```

- Assess convergence of above models *via* trace plots, autocorrelation plots and Geweke test:

```
mcmc.samples <- as.mcmc(jags.samples) # Converts samples to "MCMC object"
                                      # for diagnostics

# Traceplot and density plots for regression coefficients
# code will save to PDF in current directory.
# Execute "plot" commands only to plot to screen.
pdf("Traceplot_LogisticReg.pdf")     # Write what comes next to PDF file
plot(mcmc.samples[1][,1:4])          # For Chain 1, beta1-4
plot(mcmc.samples[1][,5:8])          # For Chain 1, beta5-7 and deviance
dev.off()                            # Stop writing to the PDF file

# Autocorrelation plots for the regression coefficients
pdf("ACF_LogisticReg.pdf")
par(omi=c(.25,.25,.25,.25))          # Create an outer margin (room for title)
autocorr.plot(mcmc.samples[1][,1:7]) # For chain 1
title("Chain 1", outer=T)            # Place title in outer margin of page

autocorr.plot(mcmc.samples[2][,1:7]) # For chain 2 (optional)
title("Chain 2", outer=T)
```

```
autocorr.plot(mcmc.samples[3][,1:7]) # For chain 3 (optional)
title("Chain 3", outer=T)
dev.off()

geweke.diag(mcmc.samples[,1:7])      # Geweke test
```

- Now change the prior on the smoking variable to reflect that you expect the prior odds ratio to be 2 (*i.e.* smokers have twice the odds of having CVD than non-smokers), while leaving the prior means on the other coefficients equal to zero and all prior variances=1000 (on the log-scale); call this model "Informative Prior 1."

```
# Informative prior 1 (Change prior mean to log(2) for b[2])
mu[2] <- log(2)

set.seed(123)
jags.samples.inform1 <- jags(data=data.list,
                             model.file=model.posterior,
                             inits=inits.model,
                             parameters.to.save=parameters.model,
                             n.iter=n.iter, n.chains=3)
print(jags.samples.inform1,digits=4)
```

- Next: on your own, estimate this same model (with the new prior mean) after increasing your conviction for this prior belief. Assert that you have 95% confidence that the OR for smoking lies between 1.5 and 2.67 (use the midpoint on the log-scale as the prior mean). Given your Normal prior, use this information to calculate the standard deviation,[2] and precision for β_2 on the log-scale. Call this model "Informative Prior 2."

```
# Informative prior 2 (Change prior precision)
sd.prior <-   # ON YOUR OWN, calculate SD for beta2 on log-scale
tau[2] <-     # ON YOUR OWN Convert sd.prior to precision

set.seed(123)
jags.samples.inform2 <- jags(data=data.list,
                             model.file=model.posterior,
                             inits=inits.model,
                             parameters.to.save=parameters.model,
                             n.iter=n.iter, n.chains=3)
print(jags.samples.inform2,digits=4)
```

(questions on next page)

[2] See lecture notes. Use a span of 2 × 1.96 for the width of the CI (don't use 4 as an approximation).

Questions

1. Using the R code provided, complete Table 1 using the posterior samples of the odds ratios. **(20 points)**

Table 1: Posterior median and 95% credible intervals for odds ratios from logistic regression model of CVD status on smoking, controlling for age, sex, and education level.

Variable	Vague prior	Informative Prior 1*	Informative Prior 2†
Current smoker (vs. non)			
Age (per year increase)			
Male sex (vs. female)			
High school education (vs. < HS)			
Some college (vs. < HS)			
College+ (vs. < HS)			

* Prior mean for OR of current smoking=2, prior variance of log-OR=1000.

† Prior mean for OR of current smoking=2, prior variance of log-OR=[fill in].

2. Using the parameterization for **Informative Prior 1**, calculate the prior 95% interval for the smoking OR. *Hint: Calculate the interval on the scale of the log-OR (β) and transform the limits.* In **one or two sentences** describe how this compares to the prior interval for **Informative Prior 2** stated in the instructions above, and **why** one may be more informative than the other. **(10 points)**

3. What seems to be more influential on the smoking effect, Informative prior 1 or Informative prior 2? In **one sentence**, briefly explain what you think is happening? **(5 points)**

4. Using the trace plots, density plots and autocorrelation plots (focus on 1st chain) from the diagnostics for the first model ("vague prior"), **briefly describe any evidence of convergence (or lack of convergence) that you see**. Attach these plots (2 pages for trace/density plots; 1 page for autocorrelation plots). **(10 points)**

5. From the results of the Geweke test, is there evidence for lack of convergence? Justify your answer. **(5 points)**