Homework: Bayesian Bias Analysis

Due 12 noon, 5 May 2020

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.**

Load required packages and read data:

```
library(R2jags)
library(epiR)
library(survival)
library(SurvRegCensCov)
library(epiR)
library(tictoc) # To time things

load("senssamp.rdata")
colnames(senssamp) <- tolower(colnames(senssamp))

# Correction to the dataset from last week's assignment:
senssamp$mortime2[senssamp$mortime2==0] <- .01 # Survival times need to be >1
```

Using a *Bayesian* approach, we will replicate part of the bias analysis described in the 2003 *Epidemiology* paper by Lash and Fink.[1] The data were obtained from the companion website to the text. See last week's assignment for additional details.

Standard analysis

Instead of a Cox proportional hazards model (which isn't straightforward in a Bayesian framework) we will use a parametric **Weibull** proportional hazards model to estimate the relationship between receiving less than definitive therapy (`defnther=1` vs. `=0`) and breast cancer-related mortality in this cohort. We will adjust models for age (categorical: `agecat1` and `agecat2`), and stage (regional vs. local: `excat1`), as we did previously.

The outcome (time to death: t) is assumed to follow a Weibull distribution, with shape parameter ρ, and rate parameter $\lambda(t|\mathbf{x}, \beta)$ where $\mathbf{x}$ is a vector of covariates (`defnther, excat1, agecat1, agecat2`):

$$t \sim \text{Weibull}(\rho, \lambda(t|\mathbf{x}, \beta)) \tag{1}$$

$$\log[\lambda(t|\mathbf{x})] = \log(\rho \times t^{\rho-1}) + \beta_1 + \mathbf{x}\beta_{2:5}. \tag{2}$$

[1] Lash, Timothy L., and Aliza K. Fink. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology* (2003): 451-458.

With this parameterization, the β coefficients represent log-hazard ratios per unit change in covariate. For the standard analysis, we assume non-informative priors on ρ and β:

$$\rho \sim \text{Log-Normal}(0, \tau = 8) \tag{3}$$

$$\beta_j \sim \text{N}(0, \tau = 0.01) \text{ for } j = 1, \ldots k. \tag{4}$$

A programming note: This analysis includes censored observations, and JAGS deals with censoring in a particular way. It requires creating a binary variable to indicate censoring (=1) or not (=0) [instead of the usual death (=1) vs. censoring (=0)], and creating **two** time variables for the event time (t) and censoring time (c):

- **For events (deaths):** t = the observed follow-up time, and c = some value $> t$ (it doesn't matter–you can set it to something greater than maximum follow-up time).
- **For censored observations (non-deaths):** t is set to missing (we don't know the actual event time), and c = the observed follow-up time (the last time we know the individual was alive). t is only known to be $> c$.
- The **censoring indicator** is assigned the `dinterval` distribution (see JAGS manual section 9.2.4 for more details).

1. Create the JAGS function to define the posterior:

```
jags.weibull <- function(){
     # SAMPLING DISTRIBUTION
     for (i in 1:N) {
          # Define the log-hazard:
          log(lambda[i]) <- b[1] + b[2]*defnther[i] + b[3]*excat1[i] +
                            b[4]*agecat1[i] + b[5]*agecat2[i];

          # Tell JAGS which observations are censored, and the distributon of
          # failure times:
          censored[i] ~ dinterval(t[i], c[i]);
          t[i] ~ dweib(shape, lambda[i]);
     }

     # Prior on betas (log-HR):
     b[1:N.x] ~ dmnorm(mu.b[1:N.x], tau.b[1:N.x, 1:N.x]) # multivariate normal prior

     # Prior on shape parameter for Weibull:
     shape ~ dlnorm(0, 8);

     # Calculate HRs:
     for (l in 1:N.x) {
          HR[l] <- exp(b[l]);
     }
}
```

2. Define parameters and data elements:

```
# Parameters and priors
N.x <- 5                        # Number of predictors in outcome model
N <- nrow(senssamp)             # Total number of observations
mu.b <- rep(0, N.x)             # Prior mean for regression parameters
tau.b <- diag(10^-2, N.x)       # Prior precision on regression parameters

# Outcome data:
# Define death (event) time and censoring time:
senssamp$t <- senssamp$c <- senssamp$mortime2
# If censored (death indicator = 0) then death time is missing
senssamp$t[senssamp$bccause == 0] <- NA
# If death (death = 1) then censoring time > death time
senssamp$c[senssamp$bccause == 1] <- 6 # Greater than all observed times
senssamp$censored <- 1 - senssamp$bccause # Indicator of censoring [non-death] (=1)

# Data and hyperparameter for JAGS
weibull.model.data <- list(N=N, N.x=N.x,
                           mu.b=mu.b, tau.b=tau.b,
                           t=senssamp$t, c=senssamp$c, censored=senssamp$censored,
                           defnther=senssamp$defnther, excat1=senssamp$excat1,
                           agecat1=senssamp$agecat1, agecat2=senssamp$agecat2)
```

3. Obtain simulations from the posterior for the above model[2] for 3 chains with 100,000 samples, thinning every 5th sample, and summarize. If you have a multi-core computer, running multiple chains in parallel will save some time (if not, you can substitute the `jags` function for `jags.parallel` below). This may take several minutes, even in parallel. Note that if you use the `jags.parallel` command you will not see a progress bar, so just let it run.

```
set.seed(123)
tic()
standard.weibull <- jags.parallel(model=jags.weibull, data=weibull.model.data,
                          parameters.to.save = c("b","HR","shape"),
                          n.iter=100000, n.thin=5, n.chains=3,
                          jags.seed=123)
toc()
print(standard.weibull)
plot(as.mcmc(standard.weibull), ask=T)
```

Execute the code above to fit this model as the standard analysis. Note the HR for the `defnther` variable (exposure), and its 95% CrI.

[2]NOTE: MCMC sampling from the PH parameterization of Weibull (shape, rate) can suffer from slow mixing (as you may note by the traceplots and autocorrelation plots). The parameterization above is fine for the purposes of this exercise, but Martyn Plummer (creator of JAGS) suggests a modification to sample from the AFT version of the Weibull instead (shape, scale). See: https://sourceforge.net/p/mcmc-jags/discussion/610036/thread/d5249e71/. Another useful alternative is a piecewise exponential model, which provides nearly identical inferences.

You may wish to compare this to a standard Weibull model from the `survival::streg` command (see R script).

Bias Analysis 1

We want to explore how robust this result is to unmeasured confounding. The model of interest would ideally include the binary unmeasured confounder (u) with the exposure (x_1) and other covariates ($x_{2:5}$):

$$\begin{aligned} t &\sim \text{Weibull}\big(\rho, \lambda\big(t|\mathbf{x}, \mathbf{u}, \beta\big)\big) \\ \log\big[\lambda\big(t|\text{x}\big)\big] &= \log\big(\rho \times t^{\rho-1}\big) + \beta_1 + \text{x}\beta_{2:5} + \beta_u u. \end{aligned} \tag{5}$$

Two *sets* of bias parameters govern the degree of confounding:

- **The relationship between u and the exposure (x_1) (and possibly other variables).**
  - We believe that the confounder u would be more common among those exposed (with less than definitive therapy) than unexposed. We will assume with 95% certainty that the prevalence of u among the exposed (p_1) ranges between (.45, .65), and that the prevalence among the unexposed (reciving definitive therapy, p_0) ranges between (.30, .45), with the mode (most likely value) at the midpoint of these distributions. We will assume these parameters follow a **Beta** distribution with shape parameters (hyperparameters) of $\big(a_0, b_0\big)$ and $\big(a_1, b_1\big)$ for p_0 and p_1, respectively:

$$p_0 \sim \text{Beta}\big(a_0, b_0\big) \tag{6}$$
$$p_1 \sim \text{Beta}\big(a_1, b_1\big). \tag{7}$$

    We will describe how to identify the hyperparameters parameters below.

  - Given p_0 and p_1, the binary unmeasured confounder u is distributed according to a Bernoulli (Binomial with 1 trial) distribution:

$$\begin{aligned} u &\sim \text{Bernoulli}\big(\pi_u\big) \\ \pi_u &= p_0\big(1 - x_1\big) + p_1 x_1 \end{aligned} \tag{8}$$

    where x_1 is the binary exposure 'defnther' indicating less than definitive therapy.

- **The relationship between u and the outcome t.**
  - We will assume that u increases the risk of death, such that *a priori*, we are 95% certain that the **hazard ratio** for u given by $\exp\big(\beta_u\big)$ in equation 5 (conditional on the other covariates) ranges between 1 and 3. We will assume a normal distribution for β_u, with the mean equal to the midpoint of this interval (on the log-scale), and the standard deviation equal to the width of this range divided by 2×1.96 (on the log sale).

1. Modify the code below in the places indicated to create a function for the posterior for JAGS to account for undertainty in the systematic error due to confounding.

```
jags.weibull.conf <- function(){
     for (i in 1:N) {
          # SAMPLING DISTRIBUTION

          # ***** MODIFY THE FOLLOWING EXPRESSION FOR THE HAZARD TO
          # ***** INCLUDE THE EFFECT OF THE UNMEASURED CONFOUNDER
          # ***** ASSUMING THE LOG-HAZARD RATIO IS CALLED b.u
          # ***** AND THE COVARIATE IS CALLED U:
          log(lambda[i]) <- b[1] + b[2]*defnther[i] + b[3]*excat1[i] +
               b[4]*agecat1[i] + b[5]*agecat2[i];

          censored[i] ~ dinterval(t[i], c[i]);
          t[i] ~ dweib(shape, lambda[i]);

          # Distributon for the unmeasured confounder:
          pi.u[i] <- p.0*(1-defnther[i]) +   # Definitive therapy
                     p.1*defnther[i]         # Less than definitive therapy

          U[i] ~ dbin(pi.u[i],1) # Sample the unmeasured confounder
     }

     # Priors on betas:
     b[1:N.x] ~ dmnorm(mu.b[1:N.x], tau.b[1:N.x, 1:N.x]) # multivariate normal prior
     b.u ~ dnorm(mu.b.u, tau.b.u) # the prior for the bias parameter b.u

     # Shape parameter for Weibull:
     shape ~ dlnorm(0,8);

     # Priors for bias parameters:
     # ***** SPECIFY PRIORS FOR THE BIAS PARAMETERS
     # ***** THAT DEFINE THE DISTRIBUTION OF U WITH RESPECT
     # ***** TO THE EXPOSURE GROUPS.
     # ***** DEFINE THESE IN TERMS OF HYPERPARAMETERS:
     # ***** a.0, b.0 (for p.U0)
     # ***** a.1, b.1 (for p.U1)
     p.0 ~  # Definitive therapy, did not die
     p.1 ~  # Less than definitive therapy, did not die

     # Calculate HRs:
     for (l in 1:N.x) {
          HR[l] <- exp(b[l]);
     }
     HR.u <- exp(b.u);
}
```

2. For the hyperparameters of the distribution of bias parameters:
   - Calculate the values for the mean (`mu.b.u`) and precision (`tau.b.u`) as described in the instructions above (assuming that $HR_u = \exp(\beta_u)$ is 95% likely to lie in a range of 1-3).
   - For the distributions of p_0 and p_1, use the function `epi.betabuster` in the `epiR` package to find the shape parameters ((a_0, b_0) and (a_1, b_1), respectively) from a Beta distribution that corresponds to the stated assumptions (refer to text above):
     - We assume that the **mode** of each prior is **at the midpoint** of each of the 95% prior intervals.
     - A 95% prior interval corresponds to a **confidence level of 97.5%** that the parameter is **greater than** the **lower limit** of the interval (e.g. for p_0 we assume that $\Pr[p_0 > .3] = 0.975$).

   Using these details, calculate the values for the `mode`, `conf` (confidence level), and lower confidence limit `x` to and input into the `epi.betabuster` function to obtain the two shape parameters (*a* and *b*) for **both** of the bias parameters on the confounder prevalence (p_0 and p_1).

```
# ***** ACCORDING TO THE ABOVE DESCRIPTION OF THE BIAS PARAMETERS:
mu.b.u <- # ***** SPECIFY THE VALUE FOR THE MEAN OF THE PRIOR FOR b.u
tau.b.u <- # ***** SPECIFY THE VALUE FOR THE PRECISION FOR THE PRIOR FOR b.u

# Uses epi.betabuster to obtain parameters for Beta distribution
# that satisfies the stated inputs (mode and lower quantile):
beta.p0 <- epi.betabuster(mode= , conf= ,
                          greaterthan = TRUE, x= )
beta.p1 <- epi.betabuster(mode= , conf= ,
                          greaterthan = TRUE, x= )

# Pulls shape parameters out of above:
a.0 <- beta.p0$shape1; b.0 <- beta.p0$shape2
a.1 <- beta.p1$shape1; b.1 <- beta.p1$shape2

# Confirm that above parameterizations yield distribution on bias
# parameters that are consistent with beliefs:
round(qlnorm(c(0.025, 0.975), mu.b.u, 1/sqrt(tau.b.u))) # Quantiles for prior HR.u
round(qbeta(c(.025, .975), a.0, b.0), 2) # Quantiles from prior for p.0
round(qbeta(c(.025, .975), a.1, b.1), 2) # Quantiles from prior for p.1

# Add these hyperparameters to the previous data list:
weibull.model.data2 <- weibull.model.data
weibull.model.data2[["mu.b.u"]] <- mu.b.u
weibull.model.data2[["tau.b.u"]] <- tau.b.u
weibull.model.data2[["a.0"]] <- a.0
weibull.model.data2[["b.0"]] <- b.0
weibull.model.data2[["a.1"]] <- a.1
weibull.model.data2[["b.1"]] <- b.1
```

```
set.seed(123)
tic()
conf.weibull <- jags.parallel(model=jags.weibull.conf, data=weibull.model.data2,
                          parameters.to.save = c("b","b.u","HR","shape","HR.u",
                                                 "p.0","p.1"),
                          n.iter=100000, n.thin=5, n.chains=3,
                          jags.seed=123)
toc()
print(conf.weibull)
plot(as.mcmc(conf.weibull), ask=T)
```

Complete and execute the code above to fit this model, referred to as Bias Analysis 1. Take note of the HR for the `defnther` variable (exposure), the HR for *u*, and the prevalence of *u* among the exposure groups, and their corresponding 95% quantile-based intervals.

Bias Analysis 2

You are interested in exploring how sensitive your bias analysis is to alternative parameterizations of the biasing relationships. Modify the code you completed above to change **only** the prior on the HR_u parameter to reflect that the assumption that the 95% prior range for HR_u lies between (0.5, 2.0). Leave all other parameters the same. (This should only require modifying and re-executing the block of code in item 2 in the previous section, although you may want to store your posterior samples in a new object, e.g. named `conf.weibull.2`.)

Modify and execute the previous code as described above to perform a second version of the bias analysis. Take note of the HR for the `defnther` variable (exposure), the HR for *u*, and the prevalence of *u* among the exposure groups, and their corresponding 95% quantile-based intervals.

Questions

1. Complete the following table with the hazard ratio and corresponding 95% quantile-based interval estimates in each cell (place the results for the standard analysis in the right most column of row 1) (**20 points**):

Table 1: Posterior medians and 95% quantile-based intervals for parameters of standard analysis and bias analysis of unmeasured confounding for relationship between less-than-definitive therapy and breast cancer mortality.

Parameter	Standard Analysis	Bias Analysis 1	Bias Analysis 2
HR less than def therapy			
HR for U	N/A		
Prevalence U in unexposed	N/A		
Prevalence U in exposed	N/A		

2. For the exposure HR (on less than definitive therapy), answer the following in a few sentences each:
   a. Assuming the distribution of the parameters in **Bias Analysis 1**, what was the direction of the bias? Offer an *intuitive* explanation based on the imposed relationship between the confounder on the outcome (death), and its relationship with the exposure. (*Hint: Consider in the context of what the the average value of the bias parameters implies.*) **(10 points**)
   b. Assuming the distribution of the parameters in **Bias Analysis 2**, what was the direction of the bias? Offer an *intuitive* explanation based on the imposed relationship between the confounder on the outcome (death), and its relationship with the exposure. (*Hint: Consider in the context of what the the average value of the bias parameters implies.*) **(10 points**)
3. For the exposure HR (on less than definitive therapy), for each of the analyses (Standard, Bias Analysis 1, Bias Analysis 2): divide the upper limit of the credible interval by the lower limit, and report the 3 credible limit ratios (CrLR). Compare these CrLRs for each of the bias analysis results to the standard analysis. Describe what you see (i.e. for each, are they more/less precise than the standard analysis). Focusing on the comparison of Bias Analysis 2 vs. Standard Analysis: why do you think the pattern is as you observe (given the average bias implied by the distribution of bias parameters)? (**10 points**)