Homework 5

Due 7 May 2019

Grading: If you provide a reasonable attempt at a FULL answer to each question then you will receive full credit for that question regardless if the answer was correct. If answers are incomplete then points will be deduced accordingly.

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.**

Bootstrap confidence interval

A little used measure of effect in epidemiologic studies is the rate advancement period (*RAP*).[1] If the rate of disease increases monotonically with age, you can express an exposure effect in terms of how much more rapidly the risk factor advances disease progression. The rate advancement period (RAP) is the ratio of the log-HR of the exposure to the log-HR of the age effect (i.e. the ratio of the β-coefficients of BMI and age from the Cox model). (There is an analogous *risk advancement period* measure–see references in footnote.)

We will estimate the RAP for the effect of obesity on cardiovascular disease in the Framingham cohort. Using the data `CVD_RAP.Rdata` and the R code provided, estimate a Cox proportional hazards model for time to incident cardiovascular disease as a function of BMI (using indicators for each of the 4-levels), age, sex, education (4-category) and current smoking status. We will assume here that log-hazard of CVD increases linearly with age.

(see next page)

[1] See Brenner H, Gefeller O, Greenland S. Risk and rate advancement periods as measures of exposure impact on the occurrence of chronic diseases. *Epidemiology*. 1993. 4(3): 229-236. and Discacciati A, Bellavia A, Orsini N, Greenland S. On the interpretation of risk and rate advancement periods. *Int J Epidemiol*. 2016. 45(1): 278-284.

```
library(foreign)
library(survival)
library(car)
load("CVD_RAP.Rdata")

comp.data$bmi_cat <- relevel(as.factor(comp.data$bmi_cat),2) # NW referent

# Estimate Cox model:
fit.cox.cvd <- coxph(Surv(timecvd, cvd)~factor(bmi_cat) + age + male +
                           factor(educ) + cursmoke,
                     data=comp.data, ties="efron")
```

The *RAP* for obesity is the ratio of the log-hazard rate for obesity (β_3) to the log-hazard rate of age (β_4):

```
# Calculate RAP:
b3 <- coef(fit.cox.cvd)[3] # Obese vs. normal effect
b4 <- coef(fit.cox.cvd)[4] # Age effect

RAP.OB <- b3/b4 # Point estimate of RAP
```

Since the *RAP* is a nonlinear combination of the model parameters we would typically use the delta method to calculate the standard error and confidence interval (which is completely appropriate).

Tasks:

- Use the code below to calculate the *RAP* and corresponding 95% confidence interval using the delta method:

```
# Confidence interval using Delta method:
RAP.OB.CI<- deltaMethod(fit.cox.cvd,"b3/b4",
                            parameterNames= paste("b",
                                  1:length(coef(fit.cox.cvd)), sep=""))
round(RAP.OB.CI$Estimate + RAP.OB.CI$SE*c(0,-1.96,1.96),2)
```

Although this is completely sensible, we could also *bootstrap* the confidence interval by resampling from the data and calculating a new *RAP* each time. This might be more appropriate if we have a small-ish sample size (the delta method is appropriate in large samples).

The procedure for bootstraping the CI is as follows:[2]

1. Construct a random sample from your dataset.
2. Estimate the *RAP* for that sample, and store the result.
3. Go back to step 1 and repeat a number of times (`N.sims`).
4. Summarize the distribution of all of the *RAP*s you calculated.

- Use the code below to calculate a bootstrapped estimate of the confidence interval for the *RAP*:

[2] See Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *International Journal of Epidemiology*. 2004; 33: 1389-1397. and Haukoos JS, Lewis RJ. Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions. *Academic Emergency Medicine*. 2005.

```
# Confidence interval via Bootstrapping:
N.sims <- 5000      # Number of simulations (resamples)

RAP.OB.BOOT <- rep(NA,N.sims)

set.seed(8765432)
for (i in 1:N.sims) {
  # Randomly sample the dataset with replacement
  data.tmp <- comp.data[sample(1:dim(comp.data)[1], replace=TRUE),]

  # Estimate Cox model on the resampled data:
  fit.cox.cvd.boot <- coxph(Surv(timecvd, cvd)~factor(bmi_cat) + age +
                                male + factor(educ) + cursmoke,
                      data=data.tmp, ties="efron")

  # Accumulate estimates of RAP
  RAP.OB.BOOT[i] <- coef(fit.cox.cvd.boot)[3]/coef(fit.cox.cvd.boot)[4]

}

# 95% quantile-based Bootstrapped confidence intervals for RAP
round(quantile(RAP.OB.BOOT, c(0.025, 0.975)),2)
mean(RAP.OB.BOOT)
median(RAP.OB.BOOT) # Probably better if distribution is skewed

# Plot the kernel density to inspect the distribution:
pdf("RAP Density.pdf")
plot(density(RAP.OB.BOOT), main="Empirical density of RAP for Obesity")
dev.off()
```

Linear regression model

We wish to estimate a regression model in a Bayesian framework for the relationshp between BMI and age, gender, smoking, and education.

Tasks:

Load required packages and read data:

```
library(R2jags)
library(coda)
require(lmtest)
require(foreign)
require(survival)
```

```
load("frmgham_recoded.Rdata")
```

- Estimate a linear regression model for BMI as a function of age, gender, smoking status, 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=$\tau = 1000^{-1}$).

```
# JAGS code for the posterior distribution:
bmi.model <- function() {
  for (i in 1:N) {
      mu.bmi[i] <- b[1] + b[2]*age[i] + b[3]*female[i] + b[4]*cursmoke[i] +
                   b[5]*educ2[i] + b[6]*educ3[i] + b[7]*educ4[i];
      bmi[i] ~ dnorm(mu.bmi[i],tau.bmi);      # Sampling distribution
    }
  tau.bmi ~ dgamma(a.tau, b.tau);             # Precision
  se.bmi <- 1/sqrt(tau.bmi);                  # Standard error of residuals

  # PRIORS ON BETAS
  b[1:Nx] ~ dmnorm(mu.b[1:Nx],tau.b[1:Nx,1:Nx]) # multivariate normal prior
}

# Extract data elements from data frame
bmi <- frmgham_recoded$bmi
age <- frmgham_recoded$age
female <- as.integer(frmgham_recoded$sex==2)
cursmoke <- frmgham_recoded$cursmoke

# Create education indicators (a shortcut using the model.matrix command)
educ1 <- model.matrix(~ -1 + factor(educ), data=frmgham_recoded)[,1]
educ2 <- model.matrix(~ -1 + factor(educ), data=frmgham_recoded)[,2]
educ3 <- model.matrix(~ -1 + factor(educ), data=frmgham_recoded)[,3]
educ4 <- model.matrix(~ -1 + factor(educ), data=frmgham_recoded)[,4]

# Constants to be passed in
N <- length(bmi);                    # Num. obs.
Nx <- 7;                             # Num. of reg. params. (w/ intercept)

# Parameters on the priors:
a.tau <- .001; b.tau <- .001;        # Parameters for prior on tau.bmi
mu.b <- rep(0,Nx);                   # Prior mean of beta
tau.b <- diag(.001,Nx);              # Prior precision

# Lists for JAGS
bmi.data <- list("N","Nx","bmi","age","female","cursmoke","educ2",
                 "educ3","educ4","mu.b","tau.b","a.tau","b.tau")

bmi.parameters <- c("b","tau.bmi", "se.bmi") # Parameters to keep track of
```

```
# bmi.inits <- function() {list (b=rep(0,Nx))}
bmi.inits <- list(list(b=rep(0,Nx)),
                  list(b=rep(-1,Nx)),
                  list(b=rep(1,Nx)))

bmi.sim <- jags(data=bmi.data,inits=bmi.inits,
              parameters.to.save=bmi.parameters,
              n.iter=10000, model.file=bmi.model,
              n.thin=1, jags.seed=110410)

print(bmi.sim,digits=4)
```

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

```
bmi.mcmc <- as.mcmc(bmi.sim)

# 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_LinearReg.pdf")     # Write what comes next to PDF file
plot(bmi.mcmc[1][,1:4])               # For beta1-4
plot(bmi.mcmc[1][,5:8])               # For beta5-6 and deviance
dev.off()                          # Stop writing to the PDF file

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

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

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

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

- Change the prior on the smoking variable to reflect that you expect the mean of the corresponding slope to be +100 (current smokers have a BMI 100 kg/m^2 greater than non-smokers [yes, this is nonsensical]), while leaving the means on the other slope parameters equal to zero and all prior variances=1000; call this model "Informative Prior 1."

```
# Informative prior 1 (Change prior mean to 1000 for beta[4])
mu.b[4] <- 100

bmi.sim.inform1 <- jags(data=bmi.data,inits=bmi.inits,
                        parameters.to.save=bmi.parameters,
```

```
                          n.iter=10000, model.file=bmi.model,
                          n.thin=1, jags.seed=110410)

print(bmi.sim.inform1,digits=4)
```

- Next, estimate this same model (with the new prior mean) after increasing your conviction for this prior belief: with a prior variance on this parameter of 0.1225 (precision of $1/0.1225$); call this model "Informative Prior 2."

```
# Informative prior 2 (Change prior precision to 1/0.1225 vor beta[4])
tau.b[4,4] <- 1/0.1225   # Gets evaluated when "jags" function is called

bmi.sim.inform2 <- jags(data=bmi.data,inits=bmi.inits,
                          parameters.to.save=bmi.parameters,
                          n.iter=10000, model.file=bmi.model,
                          n.thin=1, jags.seed=110410)

print(bmi.sim.inform2,digits=4)
```

Questions

Bootstrap confidence interval for RAP

1. Write out the statistical model that you fit to calculate the *RAP* in terms of the parameters (e.g. β-coefficients). What is the expression for the *RAP* in terms of the model parameters? **(15 points)**
2. What is the value for the *RAP* for the obese vs. normal weight exposure level and its 95% CI from the delta method? Offer an interpretation of this effect measure. **(15 points)**
3. What is the bootstrapped mean and 95% confidence interval estimate for the *RAP* for the obesity effect. How does this bootstrapped estimate of the 95% confidence interval compare to the estimate from the delta method (remembering that they are both approximations!). Turn in the density plot for the bootstrapped *RAP*. **(10 points)**

(see next page)

Linear regression model of BMI

4. Using the R code provided, complete Table 1. **(20 points)**

Table 1: Posterior means and 95% credible intervals for slope coefficient from linear regression model of BMI on age, sex, and education level.

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

* Prior mean for effect of current smoking=100, prior variance=1000.
† Prior mean for effect of current smoking=100, prior variance=0.1225.

5. What seems to be more influential, varying the prior mean, or the prior variance? In **one sentence**, briefly explain what you think is happening? **(10 points)**

6. 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). **(20 points)**

7. From the results of the Geweke test, is there evidence for lack of convergence? Justify your answer. **(10 points)**