PUBH 7440: Intro to Bayesian Analysis
Homework from Week 3 — Due Feb 15th

Stroke mortality in PA: [Insert something about how stroke is the 4th leading cause of death and thus it's super important we study trends in stroke mortality...] Here, we want to look at county-level stroke mortality rates among those aged 65–74, 75–84, and 85+ from 2016 in the state of Pennsylvania. Specifically, our data consist of the number of deaths due to stroke, Y_{ia}, from county i and age-group a out of a population of size n_{ia}, where $i = 1, \ldots, N_s$ and $a = 1, \ldots, N_g$ ($N_s = 67$ and $N_g = 3$).

1. Assume $Y_{ia} \sim \text{Pois}\left(n_{ia}\lambda_{ia}\right)$ where $\lambda_{ia} \sim \text{Gam}\left(Y_{0a}, n_{0a}\right)$. Using the pmf of the Poisson distribution and pdf of the gamma distribution below, show that this prior is a *(conditionally) conjugate prior* for λ_{ia} and write its full-conditional distribution. Note: Please use the parameterization of the gamma distribution below, *not* the one listed in Appendix A of CL3.

$$p\left(\lambda_{ia} \,|\, Y_{0a}, n_{0a}\right) = \frac{n_{0a}^{Y_{0a}}}{\Gamma\left(Y_{0a}\right)} \lambda_{ia}^{Y_{0a}-1} \exp\left[-n_{0a}\lambda_{ia}\right]$$

2. I claim that we can interpret Y_{0a} and n_{0a} as the prior number of deaths and prior population size for age a, respectively. Aside from my infallibility as your professor, why does this interpretation make sense?

3. Some of our counties have small population sizes, thus we may want to consider the use of informative priors. To do this, we want to construct our priors such that they are consistent with (a) rate estimates that we might expect and (b) prior population sizes that respect the age distribution of Pennsylvania.

   - Let $\pi_a = \sum_i n_{ia} / \sum_a \sum_i n_{ia}$ denote the proportion of Pennsylvania's population belonging to each age group.
   - Define $\boldsymbol{\lambda}_0 = (\lambda_{01}, \lambda_{02}, \lambda_{03})$ to be the vector of our prior guesses at the age-specific mortality rates — 75, 250, and 1,000 deaths per 100,000 for ages 65–74, 75–84, and 85+, respectively (e.g., $\lambda_{01} = 75/100{,}000$). These are estimated from data from 2015 across the entire US.

   If we want our priors to correspond to a county whose total 65+ population size is 10,000, explain why specifying $n_{0a} = \pi_a \times 10{,}000$ and $Y_{0a} = n_{0a}\lambda_{0a}$ could achieve this and the two above goals.

4. Due to data confidentiality issues, all counts $Y_{ia} < 10$ have been suppressed in public-use data. How can we account for this in our analysis?

5. Write out the full hierarchical model; i.e., something like:

$$p\left([\text{all of the unknown parameters}] \mid [\text{all of the } \textit{known} \text{ data}]\right) \propto [\text{Likelihood}] \times [\text{All of the priors}]\,.$$

6. Using the data on Canvas and the code outline below, write a Gibbs sampler to fit this model.

7. After obtaining samples from the posterior distribution for each λ_{ia}, obtain samples from the posterior distribution of the county-specific age-adjusted mortality rates — $\lambda_{i\cdot} = \sum_a \pi_a \lambda_{ia}$. Using the last piece of code below, create a map of the posterior medians of the age-adjusted rates. Your estimates should be *similar to* (but not the same as) the estimates on CDC WONDER.

8. **OPTIONAL**: CDC WONDER's privacy protections have issues. For instance, we can modify our request to obtain the total number of deaths in the state of Pennsylvania for each age group. While we *could* use this information to improve our imputation step, what I want you to do is:

   - Keep track of the imputed values for each iteration of the Gibbs sampler.
   - Compare the posterior distribution for the total number of deaths (i.e., the true/uncensored Y_{ia}'s plus the imputed values) to the true total death counts. Is our approach overestimating the death counts?

```
#https://wonder.cdc.gov/controller/saved/D140/D34F844
rm(list=ls())
#First we read in the data and define a few things...
stroke=read.table('2016_PA_stroke_total.txt',sep='\t',
                  stringsAsFactors=FALSE,header=TRUE)
Ng=3       #three age groups
alabs=unique(stroke$Age.Group.Code)
Ns=67     #67 counties
clabs=unique(stroke$County)

#Next we organize things a bit...
Y=array(stroke$Deaths,dim=c(Ng,Ns))
n=array(stroke$Population,dim=c(Ng,Ns))

###################
###################
#per part 4, all Y's below 10 have been suppressed
###################
thres=10      #Suppression threshold
dY=!is.na(Y)  #0 for suppressed, 1 for observed
nsupp=apply(!dY,1,sum) #how many suppressed per age
#note: we do not know the true Y's
#      CDC did the suppression, not me

###################
###################
#insert your prior info here
```

```
###################
lam0=c(75,250,1000)/100000
n0=      ###############
Y0=      ###############

###################
###################
#initialize your Gibbs sampler here
nsims=10000
lami=array(dim=c(Ng,Ns,nsims))
for(a in 1:Ng){
  lami[a,,1]=    ################
  Y[a,!dY[a,]]=  ################
  #Note: the preceding line assumes we don't care
  #      what the posterior dist of the missing
  #      Y's looks like -- I'm just plugging the
  #      current guesses directly into my data vector
}

for(it in 2:nsims){
  for(a in 1:Ng){
    ###################
    ###################
    #ADDRESS SUPPRESSED Y HERE
    ###################
    ###################

    ###################
    ###################
    #ESTIMATE LAMBDA_{ia} HERE
    ###################
    ###################
  }
}

#################
#################
#Get posterior samples
#of the age-adjusted rates
#################
aalami=array(dim=c(Ns,nsims))
for(i in 1:Ns){
  aalami[i,]=  ##################
}
```

```
##################
##################
#calculate the posterior medians
# of the age-adjusted rates
aa.med= ##################

##################
##################
#THE BELOW CODE SHOULD BE LEFT AS-IS!
#IT ASSUMES YOU NAMED
#THE POSTERIOR MEDIANS OF THE AGE-ADJUSTED RATES
#"aamed" USING THE CODE ABOVE,
#AND WILL CREATE A MAP "PAmap.png"
#THAT WILL BE SAVED TO YOUR CURRENT DIRECTORY
##################

load('penn.rdata')
install.packages(c('maptools','RColorBrewer'))
library(maptools)
library(RColorBrewer)
ncols=7
cols=brewer.pal(ncols,'RdYlBu')[ncols:1]
tcuts=quantile(aa.med*100000,1:(ncols-1)/ncols)
tcolb=array(rep(aa.med*100000,each=ncols-1) > tcuts,
            dim=c(ncols-1,Ns))
tcol =apply(tcolb,2,sum)+1

png('PAmap.png',height=520,width=1000)
par(mar=c(0,0,0,10),cex=1)
    plot(penn,col=cols[tcol],border='lightgray',lwd=.5)
    legend('right',inset=c(-.15,0),xpd=TRUE,
           legend=c(paste(
           c('Below',round(tcuts[-(ncols-1)],0),'Over'),
           c(' ',rep( ' - ',ncols-2),' '),
           c(round(tcuts,0),round(tcuts[ncols-1],0)),sep='')),
           fill=cols,title='Deaths per 100,000',bty='n',cex=1.5,
           border='lightgray')
dev.off()
```

Problem 1

- $Y_{i\alpha} \sim Pois(n_{i\alpha} \lambda_{i\alpha})$, $\lambda_{i\alpha} \sim Gamma(\gamma_{0\alpha}, \nu_{0\alpha})$

- $i \Rightarrow$ county number

  $\alpha \Rightarrow$ age group.

- $Y_{i\alpha}$ = death due to death stroke

  $n_{i\alpha}$ = population

  $\lambda_{i\alpha}$ = death rate

- $$p(Y_{i\alpha} \mid n_{i\alpha} \lambda_{i\alpha}) = \frac{e^{-(n_{i\alpha}\lambda_{i\alpha})} (n_{i\alpha}\lambda_{i\alpha})^{Y_{i\alpha}}}{(Y_{i\alpha})!}$$

- $$p(\lambda_{i\alpha} \mid \gamma_{0\alpha}, \nu_{0\alpha}) = \frac{\nu_{0\alpha}^{\gamma_{0\alpha}}}{\Gamma(\gamma_{0\alpha})} \cdot \lambda_{i\alpha}^{\gamma_{0\alpha}-1} e^{(-\nu_{0\alpha}\lambda_{i\alpha})}$$

- Posterior:

  $$p(\lambda_{i\alpha} \mid Y_{i\alpha}) \propto \frac{e^{-(n_{i\alpha}\lambda_{i\alpha})} (n_{i\alpha}\lambda_{i\alpha})^{Y_{i\alpha}}}{(Y_{i\alpha})!} \times$$

$$\times \frac{\nu_{0\alpha}^{\gamma_{0\alpha}}}{\Gamma(\gamma_{0\alpha})} \cdot \lambda_{i\alpha}^{(\gamma_{0\alpha}-1)} e^{(-\nu_{0\alpha}\lambda_{i\alpha})} \propto$$

$$\frac{e^{-(\lambda_{i\alpha})} \lambda_{i\alpha}^{y_{i\alpha}} \lambda_{i\alpha}^{\gamma_{0\alpha}-1} e^{(-\nu_{0\alpha}\lambda_{0\alpha})}}{(y_{i\alpha})!} \propto$$

$$\lambda_{i\alpha}^{(y_{i\alpha}+\gamma_{0\alpha})-1} e^{-(\nu_{0\alpha}+\nu_{i\alpha})\lambda_{i\alpha}}$$

This resembles a kernel of a gamma distribution, so, we conclude that a posterior distribution of $\lambda_{i\alpha}$ is given by

$$\lambda_{i\alpha} | y_{i\alpha} \sim \text{Gamma}(\gamma_{0\alpha}+y_{i\alpha}, \nu_{0\alpha}+\nu_{i\alpha})$$

So, a full conditional distribution can be written as

$$p(\lambda_{i\alpha} | y_{i\alpha}, \gamma_{0\alpha}, \nu_{i\alpha}, \nu_{0\alpha}) =$$

$$= \frac{(n_{o\alpha} + n_{i\alpha})^{y_{i\alpha} + y_{o\alpha}}}{\Gamma(y_{i\alpha} + y_{o\alpha})} \times \lambda_{i\alpha}^{(y_{o\alpha} + y_{i\alpha} - 1)} e^{-(n_{o\alpha} + n_{i\alpha})\lambda_{i\alpha}}$$

Problem 6

Gibbs Sampler

1) want to estimate $\lambda_{i\alpha}$ distribution.

- have some observed $Y_{i\alpha}$
  $n_{i\alpha}$

- some $Y_{i\alpha}$ need to be imputed

2) we know that $\lambda_{i\alpha} | Y_{i\alpha} \sim \text{Gamma}(Y_{0\alpha} + Y_{i\alpha},\ n_{0\alpha} + n_{i\alpha})$.

Gibbs sampler should be:

1) we have crude guesses of
75, 200, 1000 death per 100,000

2) so, $\lambda_{i1} = \dfrac{75}{100{,}000}$ for each age group α
for all counties $i = 1, 2, \ldots, 67$

just to initialize the value

3) we need to update $\lambda_{i\alpha}$

4) But!!! to update $\lambda_{i\alpha}$ we need a value of $Y_{i\alpha}$.

But!!! some of $Y_{i\alpha}$ are missing.

5) when they are missing, they are

below <u>10</u>

6) we can sample $y_{i\alpha}$ from $y_{i\alpha} \sim \text{Pois}(n_{i\alpha} \times \lambda_{i\alpha})$, and truncate <u>at 10.</u>

7) So, Gibbs sampler:

- Sample $y_{i\alpha}$, impute data, use $\lambda_{i\alpha}$
- update $n_{o\alpha} + n_{i\alpha}$, $y_{o\alpha} + y_{i\alpha}$
- use as parameters of Gamma, sample $\lambda_{i\alpha}$.