PUBH 7440: Intro to Bayesian Analysis
Take-Home Final — Due May 5th[1]...?

Suicide Deaths in MN Counties: While rates of suicide (around 15 deaths per 100,000) are far lower than rates of death due to causes like heart disease (200 deaths per 100,000) and cancer (180 deaths per 100,000), suicide is often among the leading causes of death in the U.S. Moreover, in addition to there being significant *racial* disparities in suicide rates — with rates for White Americans nearly double those for Asian and Black Americans and 40% lower than for American Indian or Alaska Natives — the *gender* disparity in suicide rates is larger than any of the other leading causes of death, with rates for men roughly four times higher than rates for women. Finally, suicide is fairly unique among other leading causes of death in the sense that suicide rates are essentially *constant* for individuals aged 15 and older (i.e., rates in 2019 were between 13.9 and 20.1 deaths per 100,000 for each of the ten-year age groups between 15–24 and 85+), so aggregating the data across multiple age groups in the name of "increasing death counts to avoid suppression on CDC WONDER" is a bit more excusable :/

With that in mind, our objective in this take-home final that you're all excited for is to look at disparities in suicide rates in Minnesota counties in 2019 by race and gender. Using data that I downloaded from CDC WONDER, we let y_{irs} denote the reported number of suicide deaths for residents of county i ($i = 1, \ldots, 87$) of race r (American Indian / Alaska Native, Asian, Black, and White) and gender s (female/male) aged 15+ out of a population of size n_{irs}. To model these data, we will assume

$$y_{irs} \sim \text{Pois}\left(n_{irs}\lambda_{irs}\right),$$

where λ_{irs} represents the suicide rate for residents of county i, race r, and gender s. Since these data are from CDC WONDER, any $y_{irs} \in [1, 9]$ are suppressed, but CDC WONDER is nice enough to tell us what the state-level totals, $y_{\cdot rs} = \sum_i y_{irs}$, are, so we'll use both the suppression criteria and these totals to impute the missing values. And since this is a Bayesian inference course, we will use the following prior distributions in our analysis:

$$\begin{aligned}
\theta_{irs} = \log \lambda_{irs} &\sim \text{Norm}\left(\beta_{0rs} + z_{irs}, \tau_{rs}^2\right) \\
\mathbf{z}_{\cdot rs} &\sim \text{CAR}\left(\sigma_{rs}^2\right) \\
\beta_{0rs} &\sim \text{Norm}\left(0, 10^4\right) \\
\tau_{rs}^2 &\sim \text{IG}\left(1, 0.01\right) \\
\sigma_{rs}^2 &\sim \text{IG}\left(1, 0.14\right).
\end{aligned}$$

Finally, I don't remember to what extent I've talked about the rationale for the priors used for τ_{rs}^2 and σ_{rs}^2, it's based on a paper by Waller et. al (1997), and while I *assume* the results wouldn't be *too* sensitive to changes in the above hyperparameters, I have nevertheless used these priors (or multivariate versions of them) in basically all of my papers :/

[1] fwiw, I think grades are due May 13th, so if you need more time, just let me know in advance...

Ground Rules

- Please refrain from talking to each other about this final (aside from maybe glowing comments about how much you're enjoying it).
- Feel free to email me with any questions/issues you may have. I might not *fully* answer your questions since this is an exam, but you're welcome to test what questions I'll answer and which I won't.
  - Relatedly, please direct all emails *to me* rather than to the TA.
- When you submit your work on Canvas, please be sure to include your code. Including it as a separate file would be ideal, but it doesn't hurt to also include it in your Word/LaTeXwrite-up, as well.

1 Preliminaries

Before you fully dig in to an analysis, it's always good to take a deep breath and think about what you're about to do and how things are going to go. With that in mind, the following questions are intended to be answered before you've ran the Gibbs sampler part of the code to fit the model (though there's obviously no way for me to enforce that).

1.1 Calculate the state-level suicide death rates (per 100,000) stratified by race and gender and comment on the racial and gender disparities in suicide rates.

1.2 Comment on the prior distributions used for the various parameters. Which priors appear to be informative and which do not (and why)?

1.3 Including the suppressed counts, y_{irs}, how many *data points* are there in our dataset? And excluding the suppressed death counts, y_{irs}, that need to be imputed, how many *parameters* are there in the model that we're going to fit? Why might that be concerning to some people? And how could you try to reassure those people that this won't destroy the world while describing the difference between fixed and random effects?

1.4 Now look at the state level death counts themselves (I'll denote them as $y_{\cdot rs}$ in my equations but they're in `Ytot` in the code). Assuming you weren't *already* having reservations about whether the analysis we're about to do is too ambitious, do you have any reservations about whether we're going to get nice stable estimates of parameters like β_{0rs}, σ^2_{rs}, and τ^2_{rs} (or, Bayes forbid, the λ_{irs} parameters)?

2 Fitting the Model

I think we can all agree that the most exciting part of Bayesian analyses is writing the code for the Gibbs sampler from scratch, so my sincere apologies that I've done that for you :/ That said, you're in luck, because *running* the code can still be an adventure :)

2.1 Specify initial values and a number of iterations (the current `nsims=500` is just a placeholder) and run the code to fit the model. I'm not really asking you a "question" here, but I suppose I'll be judging whether or not you ran enough iterations.

2.2 Make history and density plots for the β_{0rs}, σ^2_{rs}, and τ^2_{rs} parameters and comment on convergence. Before you move on to Question 2.3 below, does it *look like* there is high autocorrelation in these parameters? And, if given the choice, would you *want* to show these plots in a published paper or report or would you wait until a pesky reviewer requested them? Finally, identify an amount of burn-in that you'll use in subsequent questions.

2.3 Make autocorrelation plots for the β_{0rs} parameters. Is there a relationship between the degree of autocorrelation and the amount of data in each group (e.g., the $y_{\cdot rs}$ values)?

2.4 Make history, density, and autocorrelation plots of the estimated suicide death rates for American Indian / Alaska Native ($r = 1$) and White ($r = 4$) men ($s = 2$) in St. Louis County ($i = 69$; the county where Duluth is). Do these plots look better than the corresponding plots for β_{0rs}?

3 Bayesian Inference

As I've often alluded to in this course, if we're only interested in relatively *simple* questions — e.g., *Are there racial and gender disparities at the state level?* — we can probably get answers from simple t-tests and things like that. That said, I don't want you to only think about simple questions, so that's where the effort and creativity of Bayesian modeling comes in.

3.1 Calculate posterior medians and 95% CI for the state-level suicide death rates, stratified by race and gender using a weighted average of the λ_{irs} parameters (with weights based on the population sizes, n_{irs}). Comment on how these estimates compare to the crude rates calculated in Question 1.1 above and whether the disparities present in the crude rates appear to be "statistically significant".

3.2 Make maps of the posterior medians of the suicide death rates, stratified by race and gender and comment on the trends you see. For the purposes of comparison, try to use the same scale / color-cutoffs for each race for a given gender (i.e., use one set of cutoffs for females and another for males).

3.3 Calculate posterior medians and 95% CI for the urban/rural disparities in suicide death rates, stratified by race and gender using a weighted average of the λ_{irs} parameters (with weights based on the population sizes, n_{irs}). For the purposes of this question, assume counties whose population of residents aged 15 and older is greater than 150,000 qualify as being "urban" and all others are considered "rural"; e.g.,

```
urban=apply(n,1,sum)>150000
```

Compare these estimates to what you would obtain using the data alone (i.e., the crude rates) and comment on any differences you see and whether any of the disparities are "statistically significant".

3.4 Make maps of the posterior probability of a *racial* disparity between American Indian / Alaska Native (AI/AN), Asian, and Black men and women compared to White men and women; I'm sure this sentence was poorly worded, so I'll just say that you should end up with *six* maps (e.g., one of which will show the probability that rates for AI/AN men are greater than their White counterparts). Ideally, you'd use a *diverging* color scheme to highlight the opposite ends of the probability scale (i.e., AI/AN $>$ White vs. AI/AN $<$ White), but that's not *required.* Is there much evidence of "significant" racial disparities at the county level?

3.5 Make maps of the posterior probability of a *gender* disparity within each of the races in our data; again, this should result in *four* maps (e.g., one of which will show the probability that rates for AI/AN males are greater than their female counterparts). Ideally (again), you'd use a *diverging* color scheme to highlight the opposite ends of the probability scale (i.e., AI/AN $>$ White vs. AI/AN $<$ White), but that's not *required.* Is there much evidence of "significant" gender disparities at the county level?

4 Final Thoughts

When you're like me and almost exclusively fit big/fancy Bayesian models, you'll often find yourself wondering either (a) if you were overly ambitious when you wrote out your model and whether you should simplify things a bit or (b) how you're going to convince skeptical — often non-Bayesian — reviewers that what you did was worth doing and thus was better over the simpler alternatives. Now it's your turn :)

4.1 In Question 1.4, I asked you about whether you had any reservations about conducting a county-level analysis of these data when the state-level counts, $y_{\cdot rs}$, were pretty small for all but White men and women. After fitting the model in Section 2 and producing county-level estimates in Section 3, comment on what (if anything) you think we gained by conducting the analysis at the county level rather than at the state level.

4.2 Since these data are freely available from CDC WONDER, I *could have* obtained multiple years of data (e.g., data from 2018–2020 instead of just 2019). Keeping in mind what you said for the previous question about county- versus state-level analyses, comment on the pros and cons of *temporal* aggregation — i.e., estimating time-specific death rates, λ_{irst} versus aggregating the data over multiple years, $y_{irs} = \sum_t y_{irst}$, and estimating death rates over a period of time.

HQ’s Starter Code

```
rm(list=ls())
library(mvtnorm)
library(MCMCpack)
source(’Setup/setup_mn.r’)
tab=read.table(file=’Data/mn_suicide.txt’,sep=’\t’,header=TRUE,
               stringsAsFactors=FALSE)

##################
#Get race, gender, and county labels
##################
rlabs=unique(tab$Race)
R=length(rlabs)-1; rlabs=rlabs[1:R]
slabs=unique(tab$Gender)
S=length(slabs)-1; slabs=slabs[1:S]
clabs=unique(tab$County)
I=length(clabs)-1; clabs=clabs[1:I]

##################
#Throw out "extra" total rows
##################
tab=tab[-dim(tab)[1],] #last row is the overall total
tab=tab[tab$Gender!="",] #throw out the race totals

##################
#Set up Y, n, Ytot, Yobs
##################
Y=array(as.numeric(tab$Deaths),dim=c(I+1,S,R))
n=array(as.numeric(tab$Population),dim=c(I+1,S,R))
Ytot=Y[I+1,,]; ntot=n[I+1,,]
Y=Y[-(I+1),,]; n=n[-(I+1),,]
Yobs=apply(Y,2:3,sum,na.rm=TRUE)
nYmiss=Ytot-Yobs

###################
#Missing Y’s
###################
dY=!is.na(Y)
nsupp=apply(!dY,2:3,sum)
Ythres=c(1,9)

###################
#prior specifications
###################
```

```
gam2=10000 #beta0~N(0,gam2)
as=1; bs=1/7 #sig2~IG(as,bs)
at=1; bt=1/100 #tau2~IG(at,bt)

###################
#candidate density variances
###################
qt=array(1,dim=dim(Y))

nsims=500
set.seed(1234)
beta0=sig2=tau2=array(dim=c(S,R,nsims))
lami=z=theta=array(dim=c(I,S,R,nsims))
beta0[,,1]= #############
Ymiss=list()
for(s in 1:S){
  Ymiss[[s]]=list()
  for(r in 1:R){
    Ymiss[[s]][[r]]=array(dim=c(nsupp[s,r],nsims))
    Ymiss[[s]][[r]][,1]= #############
    Y[!dY[,s,r],s,r]=Ymiss[[s]][[r]][,1]
  }
}
sig2[,,1]= #############
tau2[,,1]= #############
z[,,,1]= #############
for(i in 1:I){
  theta[i,,,1]=beta0[,,1] + z[i,,,1]
  lami[i,,,1]=exp(theta[i,,,1])
}

for(it in 2:nsims){
for(s in 1:S){
for(r in 1:R){
    ############
    #Update Y; account for state total and bounds
    if(nsupp[s,r]>0){
      nlam=n[!dY[,s,r],s,r]*lami[!dY[,s,r],s,r,it-1]
      good=FALSE; attempt=0
      while(!good){
      remain=nYmiss[s,r]

if(attempt%%2==0){
      Yord=order(nlam)
      for(eye in 1:(nsupp[s,r]-1)){
        pie=nlam[Yord[eye]]/sum(nlam[Yord[eye:nsupp[s,r]]])
        probs=dbinom(Ythres[1]:Ythres[2],remain,pie)
        Ymiss[[s]][[r]][Yord[eye],it]=ifelse(remain==0,
                                              0,
                                              sample(Ythres[1]:Ythres[2],1,prob=probs))
        remain=remain-Ymiss[[s]][[r]][Yord[eye],it]
      }
      Ymiss[[s]][[r]][Yord[nsupp[s,r]],it]=remain
}else{
      Yord=1:nsupp[s,r]
      for(eye in 1:(nsupp[s,r]-1)){
        pie=nlam[eye]/sum(nlam[eye:nsupp[s,r]])
        probs=dbinom(Ythres[1]:Ythres[2],remain,pie)
        Ymiss[[s]][[r]][eye,it]=ifelse(remain==0,
                                       0,
                                       sample(Ythres[1]:Ythres[2],1,prob=probs))
        remain=remain-Ymiss[[s]][[r]][eye,it]
      }
      Ymiss[[s]][[r]][nsupp[s,r],it]=remain
}
      good=min(Ymiss[[s]][[r]][,it])>=Ythres[1] &
           max(Ymiss[[s]][[r]][,it])<=Ythres[2]
      attempt=attempt+1
      if(!good & attempt>=10){cat('attempt',attempt,'\n')}
      }
      Y[!dY[,s,r],s,r]=Ymiss[[s]][[r]][,it]
    }

    ############
    #update beta0
    bvar=(I/tau2[s,r,it-1] + 1/gam2)^(-1)
    bmean=bvar * (sum(theta[,s,r,it-1]-z[,s,r,it-1])/tau2[s,r,it-1])
    beta0[s,r,it]=rnorm(1,bmean,sqrt(bvar))

    ############
    #update z
    z[,s,r,it]=z[,s,r,it-1]
    for(i in 1:I){
      mui=mean(z[neigh[[i]],s,r,it])
      sigi=sig2[s,r,it-1]/m[i]

      zvar=1/(1/tau2[s,r,it-1] + 1/sigi)
      zmean=zvar * ( (theta[i,s,r,it-1]-beta0[s,r,it])/tau2[s,r,it-1] + mui/sigi )
      z[i,s,r,it]=rnorm(1,zmean,sqrt(zvar))
    }
    z[,s,r,it]=z[,s,r,it]-mean(z[,s,r,it])  #sum-to-zero constraint

    ############
    #update theta
    for(i in 1:I){
      ts=rnorm(1,theta[i,s,r,it-1],qt[i,s,r])
      ra=Y[i,s,r] * (ts - theta[i,s,r,it-1])
      rb=n[i,s,r] * (exp(ts) - exp(theta[i,s,r,it-1]))
      rc=((ts-beta0[s,r,it]-z[i,s,r,it])^2 -
          (theta[i,s,r,it-1]-beta0[s,r,it]-z[i,s,r,it])^2)
      r0=exp(ra - rb - rc/(2*tau2[s,r,it-1]))
      theta[i,s,r,it]=ifelse(r0>runif(1),ts,theta[i,s,r,it-1])

      lami[i,s,r,it]=exp(theta[i,s,r,it])
    }

    ############
    #update tau2
    tau2[s,r,it]=1/rgamma(1,I/2 + at,
                 sum((theta[,s,r,it]-beta0[s,r,it]-z[,s,r,it])^2)/2 + bt)

    ############
    #update sig2
    bss=0
    for(i in 1:I){
      bss=bss+z[i,s,r,it]^2*m[i] - z[i,s,r,it]*sum(z[neigh[[i]],s,r,it])
    }
    sig2[s,r,it]=1/rgamma(1,(I-1)/2 + as, bss/2 + bs)
  }
  }
  if(it/100 == floor(it/100)){
    cat('Iteration:',it,'\n')
    for(s in 1:S){
    for(r in 1:R){
    for(i in 1:I){
      trate=mean(theta[i,s,r,it-100+1:99]!=theta[i,s,r,it-100+2:100])
      trate=ifelse(trate>.75,.75,ifelse(trate<.2,.2,trate))
      qt[i,s,r]=qt[i,s,r]*trate/0.44
    }
    }
    }
  }
}
```