PUBH 7440: Intro to Bayesian Analysis
Midterm (Take-Home Portion) — Due March 14

Incidence of low weight births in PA: [Insert text saying why looking at the incidence of low birth weight is important]. Here, we let y_{ir} denote the number of low weight births from mothers of race r ($r = 1$ white, $r = 2$ black) in county i out of a total of n_{ir} births. To model these data, we will assume:

$$y_{ir} \sim \text{Bin}\left(n_{ir}, \pi_{ir}\right), \text{ where } \text{logit}\left(\pi_{ir}\right) = \theta_{ir} \sim \text{Norm}\left(\beta_{0r}, \sigma_r^2\right),$$

and where π_{ir} represents the incidence rate. Assuming standard priors for $\beta_{0r} \sim \text{Norm}\left(0, \tau^2\right)$ and $\sigma_r^2 \sim \text{IG}\left(0.001, 0.001\right)$, with $\tau^2 = 10{,}000$, answer the following questions:

1. Write the full hierarchical model.

$$p\left(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma} \,|\, \mathbf{y}\right) \propto \prod_{i,r} \left[\text{Bin}\left(y_{ir} \,|\, n_{ir}, \pi_{ir}\right) \times \text{LogitNorm}\left(\pi_{ir} \,|\, \beta_{0r}, \sigma_r^2\right)\right] \prod_{r} \left[\text{Norm}\left(\beta_{0r} \,|\, 0, \tau^2\right) \times \text{IG}\left(\sigma_r^2 \,|\, a, b\right)\right]$$

   Note that you could have replaced $\boldsymbol{\pi}$ with $\boldsymbol{\theta}$ and/or wrote that $\theta_{ir} \sim \text{Norm}\left(\beta_{0r}, \sigma_r^2\right)$ instead of the logit-normal expression for π_{ir} above.

2. Derive the full-conditional distributions for β_{0r}, π_{ir}, and σ_r^2. Which parameters have full-conditional distributions we can sample from directly, and which parameters require Metropolis steps to sample?

   Full-conditional distributions for β_{0r} and σ_r^2 are the same as they were in HW4 (and both can be sampled directly from their full-conditionals), so I'll just focus on the update for θ_{ir}.

   Before we start discussing our updates for everything, let's recall that if $y_{ir} \sim \text{Bin}\left(n_{ir}, \pi_{ir}\right)$ and $\text{logit}\left(\pi_{ir}\right) = \theta_{ir}$, then the pmf for y_{ir} looks like:

$$\begin{aligned} p\left(y_{ir} \,|\, \beta_0, \beta_1, z_i\right) &= \binom{n_{ir}}{y_{ir}} \pi_{ij}^{y_{ij}} \times \left(1 - \pi_{ij}\right)^{n_{ir} - y_{ij}} \\ &\propto \left(\frac{e^{\theta_{ir}}}{1 + e^{\theta_{ir}}}\right)^{y_{ij}} \times \left(\frac{1}{1 + e^{\theta_{ir}}}\right)^{n_{ir} - y_{ij}} \\ &\propto e^{y_{ir}\theta_{ir}} \times \left(\frac{1}{1 + e^{\theta_{ir}}}\right)^{n_{ir}}. \end{aligned} \tag{1}$$

   Since $\theta_{ir} \in (-\infty, \infty)$, we can sample it from a normal — and thus *symmetric* —

candidate density. As a result, our Metropolis acceptance ratio can be expressed as:

$$
\begin{aligned}
r =& \frac{e^{y_{ir}\theta^*_{ir}}}{e^{y_{ir}\theta^{(\ell)}_{ir}}} \times \left[\frac{1+e^{\theta^{(\ell)}_{ir}}}{1+e^{\theta^*_{ir}}}\right]^{n_{ir}} \times \frac{\exp\left[-\frac{\left(\theta^*_{ir}-\beta_{0r}\right)^2}{2\sigma_r^2}\right]}{\exp\left[-\frac{\left(\theta^{(\ell)}_{ir}-\beta_{0r}\right)^2}{2\sigma_r^2}\right]} \\
=& \exp\left[y_{ir}\left(\theta^*_{ir}-\theta^{(\ell)}_{ir}\right)\right] \times \left[\frac{1+e^{\theta^{(\ell)}_{ir}}}{1+e^{\theta^*_{ir}}}\right]^{n_{ir}} \times \exp\left[-\frac{(\theta^*_{ir}-\beta_{0r})^2-\left(\theta^{(\ell)}_{ir}-\beta_{0r}\right)^2}{2\sigma_r^2}\right].
\end{aligned}
$$

When coding this up, we should do as much of this on the log-scale as possible because of the potential for large values of y_{ir} and n_{ir}.

```
for(i in 1:Ns){
  ts=rnorm(1,theta[i,a,it-1],qt[i,a])
  ra=Y[i,a] * (ts - theta[i,a,it-1])
  rb=(1+exp(theta[i,a,it-1]))/(1+exp(ts))
  rc=((ts-beta0[a,it])^2 - (theta[i,a,it-1]-beta0[a,it])^2)
  r=exp(ra + n[i,a]*log(rb) - rc/(2*sig2[a,it-1]))
  theta[i,a,it]=ifelse(r>runif(1),ts,theta[i,a,it-1])
}
```

3. Write code to fit the model, and use $\beta_{0r} = 0$ and $\sigma_r^2 = 1$ as initial values.

   - Make history plots of β_{0r} and σ_r^2 for both races and assess model convergence. Is burn-in required? If so, how much?

   Full code is below. When initializing $\beta_{0r} = 0$ and $\sigma_r^2 = 1$, *some amount* of burn-in is certainly required. In my code, I didn't end up needing *that much* burn-in because the θ_{ir} parameters didn't ultimately have that far to go, but I nevertheless decided to discard the first 2,000 iterations as burn-in.

4. Suppose we're interested in investigating racial disparities in the incidence of low weight births. Using the β_{0r} terms, make a histogram of the posterior distribution of the log odds ratio. Does this indicate evidence of a "significant" racial disparity? (Hint: The log odds ratio is represented by γ_1 in the conventional regression model parameterization, $E\left[\theta_{ir} \,|\, \boldsymbol{\gamma}, \sigma_r^2\right] = \gamma_0 + \gamma_1 \times (r-1)$ where $r = 1, 2$, so you'll need to first write γ_1 as a function of the β_{0r} parameters.)

   I struggled with how to *write* this question, but the solution is meant to be relatively simple:

$$
\begin{aligned}
\beta_{0;1} =& \gamma_0 + \gamma_1 \times (1-1) = \gamma_0 \\
\beta_{0;2} =& \gamma_0 + \gamma_1 \times (2-1) = \gamma_0 + \gamma_1,
\end{aligned}
$$

and thus $\beta_{0;2} - \beta_{0;1} = \gamma_1$ corresponds to the log odds ratio. As shown in the code below, our estimate of $\gamma_1 = 0.72$ with a 95% CI of (0.64, 0.81), so this is (a) *highly* significant and (b) indicative of higher incidence of low birth weight for black mothers.

5. Now suppose we're interested in *geographic* trends in the incidence of low weight births by race and in their racial disparities. Using the mapping code from HW3/HW4, make the following maps:

   - The incidence of low weight births for white mothers.
   - The incidence of low weight births for black mothers.
   - The black/white ratio of the incidence of low weight births.

   See maps made in code...

6. Finally, make histograms of posterior distribution of the black/white ratio of the incidence of low weight births in Philadelphia County ($i = 51$) and Sullivan County ($i = 57$) and compare these to their respective crude estimates (i.e., the ratio of the crude incidence rates, y_{ir}/n_{ir}, for black and white mothers in both counties) and the statewide averages (i.e, the ratio of $\sum_i y_{ir} / \sum_i n_{ir}$ for black and white mothers). Are the posterior distributions consistent with either/both of these estimates based on the data? From a statistical perspective, would you have any reservations about presenting these results?

   Let's treat these two counties separately:

   - **Philadelphia County**: As far as I'm concerned, it looks like the posterior distribution is consistent with the racial disparities at both the state and local level. This is likely due (in part) to the fact that Philadelphia is the largest county in the state and thus is a primary driver of the rates (and disparities in rates) at the state level.
   - **Sullivan County**: There were only *TWO* low-weight births to white mothers in Sullivan County and *ZERO* *births* to black mothers, so I don't trust (and wouldn't want to report) any rate estimates from this county, much less be comfortable declaring that there is a *statistically significant* racial disparity.

   At the end of the code below, I've included another one of those "relative precision" plots (as discussed in relation to HW4), which suggests that the model is contributing roughly 50 low-weight births of each race to each county. While this might not overpower the data from *white* mothers in many counties, it certainly overpowers the data from *black* mothers in the vast majority of counties.

1 R Code

```
rm(list=ls())
#First we read in the data and define a few things...
load(file='midterm_data.rdata')
list2env(mdata,envir=sys.frame(sys.nframe()))

##################
#Y is a 67x2 array of the number of low-weight births
#  in PA counties by White mothers vs. Black mothers
hist(Y)
##################
#n is a 67x2 array of the number of total births
#  in PA counties by White mothers vs. Black mothers
hist(n)

Ns=dim(Y)[1] #67 counties
Ng=dim(Y)[2] #2 races being considered

####################
#prior specifications
####################
tau2=10000  #beta0~N(0,tau2)
as=bs=0.001   #sig2~IG(as,bs)

####################
#candidate density variances
####################
qt=array(1,dim=c(Ns,Ng))

nsims=10000
beta0=sig2=array(dim=c(Ng,nsims))
theta=array(dim=c(Ns,Ng,nsims))
for(a in 1:Ng){
  beta0[a,1]= 0 #prior mean might not be wise here...
#  beta0[a,1]=log( sum(Y[a,dY[a,]])/sum(n[a,dY[a,]]) )
  sig2[a,1]=1  #bs/as = 1, so this seems neutral
  for(i in 1:Ns){
    theta[i,a,1]=beta0[a,1] #0
  }
}

for(it in 2:nsims){
  for(a in 1:Ng){
```

```
    ############
    #update beta0
    bvar=(Ns/sig2[a,it-1] + 1/tau2)^(-1)
    bmean=bvar * (sum(theta[,a,it-1])/sig2[a,it-1])
    beta0[a,it]=rnorm(1,bmean,sqrt(bvar))

    ############
    #update theta
    for(i in 1:Ns){
      ts=rnorm(1,theta[i,a,it-1],qt[i,a])
      ra=Y[i,a] * (ts - theta[i,a,it-1])
      rb=(1+exp(theta[i,a,it-1]))/(1+exp(ts))
      rc=((ts-beta0[a,it])^2 - (theta[i,a,it-1]-beta0[a,it])^2)
      r=exp(ra + n[i,a]*log(rb) - rc/(2*sig2[a,it-1]))
      theta[i,a,it]=ifelse(r>runif(1),ts,theta[i,a,it-1])
    }

    ############
    #update sig2
    sig2[a,it]=1/rgamma(1,Ns/2 + as, sum((theta[,a,it]-beta0[a,it])^2)/2 + bs)
  }
  if(it/100 == floor(it/100)){
    cat('Iteration:',it,'\n')
    for(a in 1:Ng){
      for(i in 1:Ns){
        trate=mean(theta[i,a,it-100+1:99]!=theta[i,a,it-100+2:100])
        trate=ifelse(trate>.75,.75,ifelse(trate<.2,.2,trate))
        qt[i,a]=qt[i,a]*trate/0.44
      }
    }
  }
}

##################
#Problem 3
##################
par(mfrow=c(1,2))
matplot(t(beta0),type='l')
matplot(t(sig2),type='l')
cat("Some burn-in is required;
I'm going to use 2,000 because that's what I used before...\n")
burnin=1:2000

##################
#Problem 4
```

```
##################
logOR=beta0[2,]-beta0[1,]
par(mfrow=c(1,1))
hist(logOR[-burnin],breaks=100)
cat("The entire histogram above is greater than 0,
so we have strong evidence of a racial disparity\n")

##################
#Problem 5
##################
pii=exp(theta)/(1+exp(theta))
pci=apply(pii[,,-burnin],1:2,quantile,c(.5,.025,.975))

load('penn.rdata')
#install.packages(c('maptools','RColorBrewer'))
library(maptools)
library(RColorBrewer)
ncols=7
cols=brewer.pal(ncols,'RdYlBu')[ncols:1]

for(a in 1:Ng){
tcuts=quantile(pci[1,,a],1:(ncols-1)/ncols)*100
tcolb=array(rep(pci[1,,a]*100,each=ncols-1) > tcuts,
            dim=c(ncols-1,Ns))
tcol =apply(tcolb,2,sum)+1

png(paste('PAmap_',a,'.png',sep=''),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)],2),'Over'),
           c(' ',rep( ' - ',ncols-2),' '),
           c(round(tcuts,2),round(tcuts[ncols-1],2)),sep='')),
           fill=cols,title='LW Births per 100',bty='n',cex=1.5,
           border='lightgray')
dev.off()
}

rat=pii[,2,]/pii[,1,]
rci=apply(rat[,-burnin],1,quantile,c(.5,.025,.975))
tcuts=quantile(rci[1,],1:(ncols-1)/ncols)
tcolb=array(rep(rci[1,],each=ncols-1) > tcuts,
            dim=c(ncols-1,Ns))
tcol =apply(tcolb,2,sum)+1
```

```
png('PAmap_BW_disparity.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)],2),'Over'),
           c(' ',rep( ' - ',ncols-2),' '),
           c(round(tcuts,2),round(tcuts[ncols-1],2)),sep='')),
           fill=cols,title='B/W Disparity',bty='n',cex=1.5,
           border='lightgray')
dev.off()

###################
#Problem 6
###################
par(mfrow=c(1,2))
for(i in c(51,57)){
  hist(rat[i,-burnin],breaks=100,main=paste(penn$NAME[i],'County'),
       xlab='B/W Disparity')
  abline(v=(Y[i,2]/n[i,2])/(Y[i,1]/n[i,1]),col=2)
  abline(v=(sum(Y[,2])/sum(n[,2]))/(sum(Y[,1])/sum(n[,1])),col=4)
}

###################
#Bonus / Model Informativeness
###################
b.med=apply(beta0[,-burnin],1,median)
p0=exp(b.med)/(1+exp(b.med))
par(mfrow=c(1,2))
for(k in 1:Ng){
plot(pci[1,,k]/(pci[3,,k]-pci[2,,k]),x=Y[,k])
alpha=0.5; a=b=0
curve(qbeta(.5,x+a,x*(1/p0[k]-1)+b)/
     (qbeta(.975,x+a,x*(1/p0[k]-1)+b)-
      qbeta(.025,x+a,x*(1/p0[k]-1)+b)),
        from=1,to=max(Y[,k]),col=2,lty=1,add=TRUE)
}
```