Bayesian modeling of state wise Tuberculosis (TB) prevalence in the United States

Student Name: Ibur Rahman

Supervisor: Mahbubul Majumder, PhD

Abstract

Despite strong monitoring and surveillance, the United States is not free from Tuberculosis (TB) burden. The state-wise TB data collected from TBFACTS.ORG for three consecutive years (2013-2015). We mainly focus on 2015 data. Some exploratory analysis was done to show the state wise and year wise comparison of number of TB cases. There exists remarkable variation among TB rates across states. Usually observed estimate of state's TB rates with lower population provides poor estimates. Our analysis conforms that Bayesian exchangeable model provides better estimates.

Introduction

Inspite of improvement in Tuberculosis (TB) eradication over the years, the United States is still not free from TB burden as well as from deaths caused by TB. In 1953, per 100,000 people, the TB prevalence was recorded as 52.6 and the death rate was 12.4 due to TB. In contrast, in 2014, the TB prevalence per 100,000 was 2.9 and the death rate (per 100,000) caused by TB was 0.2 only. However, the data record illustrate that the TB prevalence across states encounter disproportionately. This study intends to estimate the state-wise true TB prevalence by applying Bayesian Hierarchical model technique.

Method

Data on state-wise population and observed number of TB cases (combined both US born and non-US born) for the year, 2013-2015 have been extracted from TBFACTS.ORG [1]. As a part of exploratory analysis, we illustrate some graphical presentations. We mainly focus on 2015 cases. We explore intensity-wise spatial distribution of TB prevalence across states for the year 2015. With the assumption of Poisson TB cases, Bayesian model with prior of equal rates as well as with prior belief of exchangeability concept [2] fitted to the data.In subsequent steps, we perform formulation of posterior distribution as well as simulation from posterior, posterior inference, comparisons of states based on posterior TB rates, Bayesian sensitivity analysis, and finally posterior predictive model checking. The analysis is done by using R packages, in particular by using "LeranBayes" and Tableau.

We denote the state-wise population and TB cases by z_i and x_i respectively for the 51 states in the United States. We consider that each state has true TB rate λ_i and we wish to estimate those rates. One option is simply estimate the true rates by using individual state TB rates as $\frac{x_1}{z_1}$, $\frac{x_2}{z_2}\ldots$ $\frac{x_5 1}{z_5 1}$. Unfortunately, these individual estimates could be poor estimates especially in case small exposures(states with small population size).In this situation, one can assume that true TB rates are equal across states and to be estimated by (1).However, this pooled estimate is questionable, as one would expect some variation in the true rates.A third compromise would be combining the individual and polled estimated by using (2) where $0 < \lambda < 1$ the shrinkage parameter.This estimate shrinks or moves individual estimates $\frac{x_i}{z_i}$ towards pooled estimate. In Bayesian setting, it is plausible to believe a priori that the true rates are similar in size, which postulates dependence structure parameters.Let $x_i \sim Poission(z_i\lambda)$, i=1,..51 and common TB rate λ is assigned as standard non-informative prior as (3), then the posterior density is given by (4), which is a gamma density

with parameters $\sum_{i=1}^{51} x_i$ and $\sum_{i=1}^{51} z_i$. One of the general method for Bayesian model suitability checking is to fit posterior predictive distribution given by (5), where x_i^* denote the number of TB cases in state i with population z_i in a future sample or time. The first term in the right hand side of (5) is Poisson density and the second term is the posterior density.We check the consistency of observed number of TB cases for the particular state x_i by using (6), where we compute the probability that the future observation x_i^* is at least as extreme as x_i.Since modeling with equal rates depicts poor model fitting, we move forward with prior belief of exchangeability.In this stage, we assume that true TB rates is random sample from a gamma distribution of the form (7) with parameters (α, $\frac{\alpha}{\mu}$, and we use subjective prior of μ and α of the form (8) and (9) respectively. Usually z_0 is specified as the median of α. In our analysis we used $z_0 = 0.53$ subjectively.The posterior distribution of true rates λ is $gamma(x_i + \alpha, z_i + \frac{\alpha}{\mu}$. Then the conditional mean of λ is given by (10), which is a form of shrinkage estimator.

The hyper- parameters α and μ to be estimated from the marginal posterior as (11). With the transformation of parameters in (12) and simulation from the marginal posterior in (13), we construct the contour plot and checked the performance of the simulation from posterior. After generating the hyper-parameter μ and α and simulated sample for true rates from the joint posterior distribution we perform various inferences. The first one is examining the shrinkage toward individual observed or pooled estimates. The posterior mean of true TB rate for state i can be written as (14), where $B_i = \frac{\alpha}{(\alpha + z_i\mu)}$ is the shrinkage if the i^{th} state TB rate. We plot mean shrinkage against logarithm of i^{th} state population. We compute posterior mean of true rate for the state by (15) and compared best state (lowest TB rate) in terms of probability. We perform sensitivity of our estimates by using (16) and (17). Finally, we performed posterior predictive model checking for our latest fitted model.

$\frac{\sum_{i=1}^{51} x_i}{\sum_{i=1}^{51} z_i}$ ..... (1)

$(1-\lambda)\frac{x_i}{z_i} + \lambda\frac{\sum_{i=1}^{51} x_i}{\sum_{j=1}^{51} z_i}$ ..... (2)

$g(\lambda) \propto \frac{1}{\lambda}$ ..... (3)

$g(\lambda|data) \propto \frac{1}{\lambda}\prod_{i=1}^{51}[\lambda^{x_i} exp(-z_i\lambda)] = \lambda^{\sum_{i=1}^{51} x_i - 1} exp(-\sum_{i=1}^{51} z_i\lambda)$ ..... (4)

$f(x_i^*|z_i, x) = \int fP(x_i^*|z_i\lambda)g(\lambda|x)d\lambda$ .....(5)

$min(P(x_i^* \leq x_i), P(x_i^* \geq x_i))$ ..... (6)

$g(\lambda|\alpha, \mu) = \frac{(\alpha/\mu)^\alpha exp(-\alpha\lambda|\mu)}{\Gamma(\alpha)}, \lambda > 0$ ..... (7)

$g(\lambda_1, \lambda_2|\alpha_0) \propto \frac{(\lambda_1\lambda_2)^{(\alpha_0 - 1)}}{(\alpha_0(\lambda_1+\lambda_2)+b)^{(2\alpha_0+\alpha)}}$ $g(\mu) \propto \frac{1}{\mu}, \mu > 0$..... (8)

$g(\alpha) = \frac{z_0}{(\alpha+z_0)^2}, \alpha > 0$ .....(9)

$E(\lambda_i|x, \alpha, \mu) = \frac{x_i+\alpha}{z_i+\alpha|\mu} = (1-B_i)\frac{x_i}{z_i} + B_i\mu, B_i = \frac{\alpha}{\alpha+z_i\mu}$..... (10)

$P(\alpha, \mu|data) = K\frac{1}{\Gamma^{51}(\alpha)}\prod_{i=1}^{51}[\frac{(\alpha|\mu)^\alpha\Gamma(\alpha+x_i)}{(\alpha|\mu+z_i)^{(\alpha+x_i)}}]\frac{z_0}{(\alpha+z_0)^2}\frac{1}{\mu}$ .....(11)

$\theta_1 = log(\alpha), \theta_2 = log(\mu)$ ..... (12)

$p(\theta_1, \theta_2|data) = K\frac{1}{\Gamma^{51}(\alpha)}\prod_{j=1}^{51}[\frac{(\alpha|\mu)^\alpha\Gamma(\alpha+x_i)}{(\alpha|\mu+z_i)^{(\alpha+x_i)}}]\frac{z_0\alpha}{(\alpha+z_0)^2}$ .....(13)

$E(\lambda|data) \approx (1 - E(B_i|data))\frac{x_i}{z_i} + E(B_i|data)\frac{\sum_{i=1}^{51} x_i}{\sum_{i=1}^{51} z_i}$..... (14)

$E\frac{(x_i+\alpha)}{(z_i+\alpha|\mu)}$ ..... (15)

$g(\theta_1|z_0) = \frac{z_0 exp(\theta_1)}{(z_0+exp(\theta_1))^2}$ ..... (16)

$w(\theta_1) = \frac{g(\theta_1|z_0=5)}{g(\theta_1|z_0=0.53)}$ .....(17)

Results

The following figure (1) illustrates that the national TB statistics United State reported that in 2015 TB cases was 9560 compared with 9565 and 9410 in 2013 and 2014 respectively.In 2014, overall TB cases reduced by 155 as compared to 2013. In 2015,the same increased by 150 as compared to 2015.

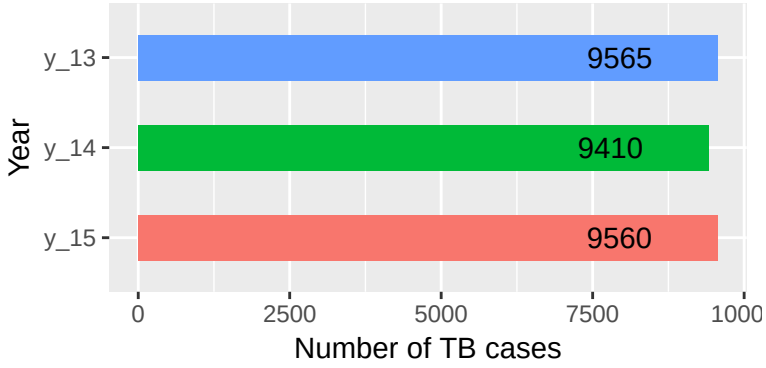

Figure 1: Number of TB cases in the year 2013-2015

In the following figure (2),State-wise TB cases are different with US born and non US born people.The figure (2) illustrates that in the most of the states TB cases of foreign born people were higher than the US born people.In California, New York,Texas,Florida,Illinois, New Jersey foreign born TB cases were larger than the US born TB cases,whereas in Alabama, Alaska, Arizona,Mississippi,South Carolina,Louisiana US born TB cases were larger than the foreign born TB cases.Overall,The highest number of TB cases was appeared in California,whereas Wyoming was the lowest in 2015.

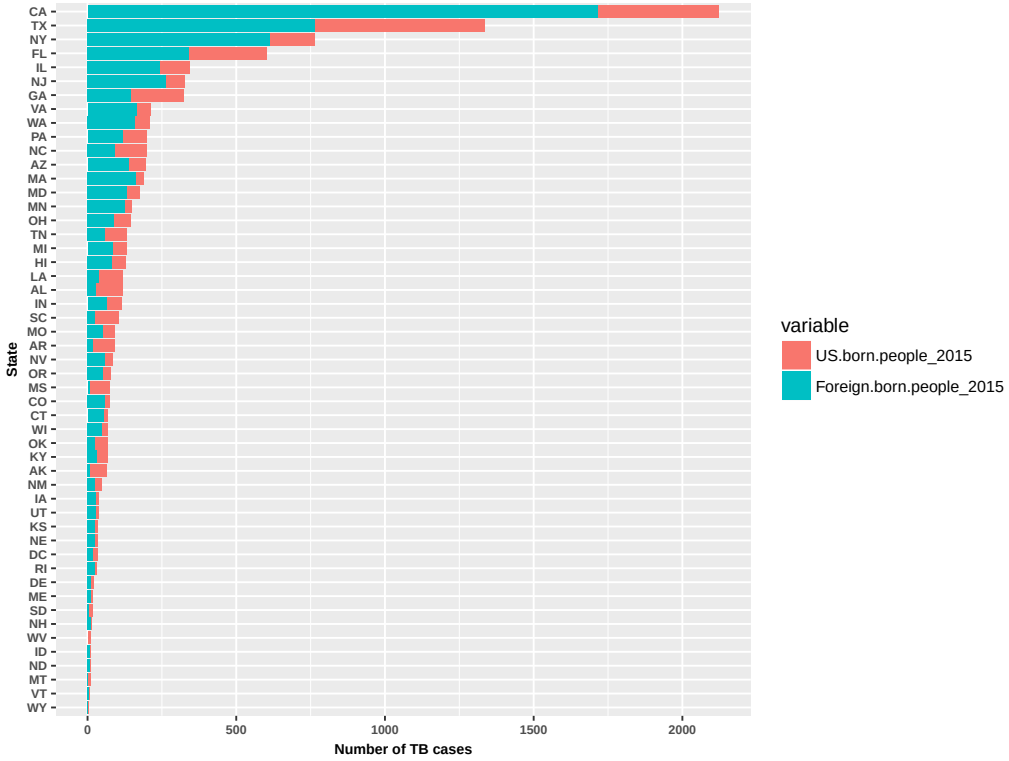

Figure 2: State-wise number of TB cases comparision for US born and foreign born in 2015

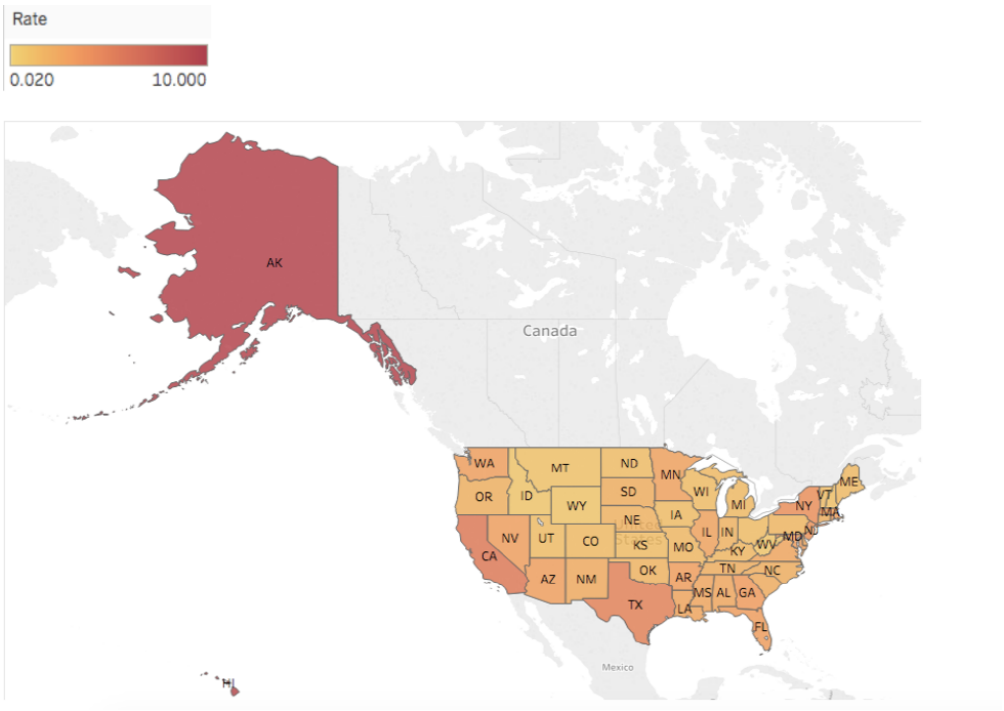

Figure 3: TB rates per 100,000 people in 2015

The figure(3) illustrates that the TB rate per 100,000 through out the states.In this map,darker color represent larger quantity, and lighter color represents smaller quantity.Hence, Darker and lighter color means the higher rate and lower rate of TB cases.In Alaska, Hawaii ,California , New York ,Texas,Florida had higher TB prevalence rate , whereas Wyoming,Idaho,West Virginia, Montana had lower TB prevalence rate.

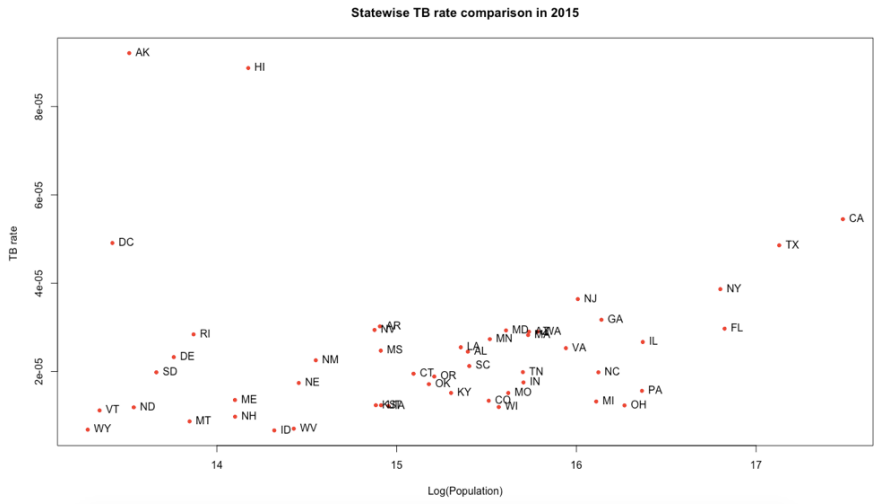

Figure 4: Statewise TB rate comparison in 2015

Figure (4) illustrates the state-wise observed TB rate in relation to logarithm of the population of the respective states. Some of the states (WY, MT, ID, and WV, etc. ;) with lower population demonstrate low TB rate while some of the states(DC, AK, HI) experience high TB rates. On the other hand, the states TX, CA with higher population size experience middle type TB rates.The observed estimates of the rates

with lower population and lower observed TB cases states might be poor estimates, as they are close zero. Therefore, we plan to estimate the true rates in Bayesian setting. As the first step in fitting Bayesian model, we assume true equal TB rates across states we obtain the posterior distribution of the rates as (4).

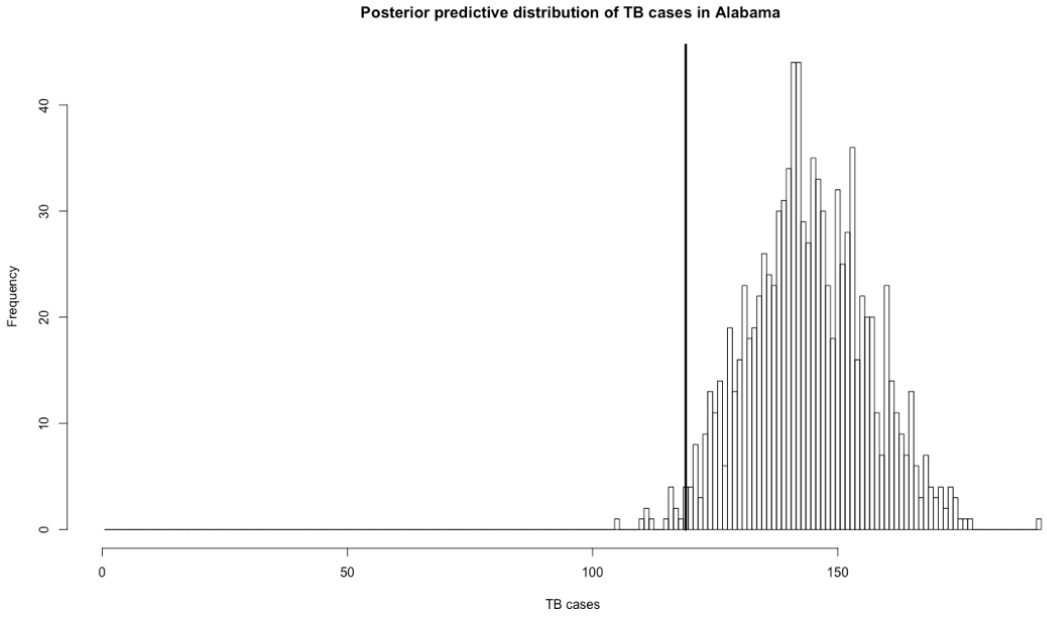

Figure 5: Posterior predictive distribution of TB cases in Alabama

To check the model fitness we produce posterior predictive distribution, simulate the sample (1000 times) from the distribution and graph (histogram) them as well figure (5).The solid (black) vertical line represents the observed number of TB cases for the Alabama state and it located at the left end of the histogram.

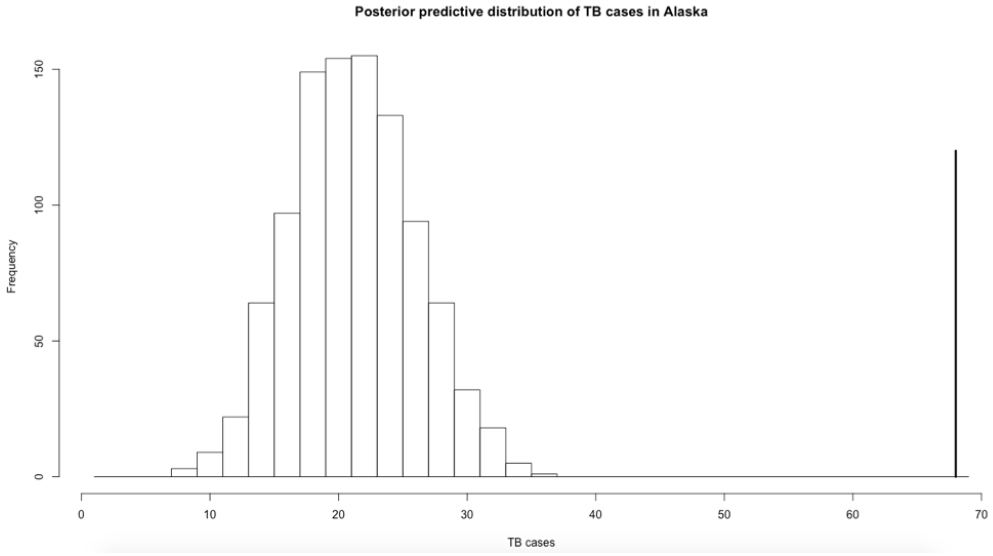

Figure 6: Posterior predictive distribution of TB cases in Alaska

We also simulate number of TB cases for the state Alaska from the posterior predictive distribution and construct histogram in figure (6). In this case, the observed number of TB cases in Alaska located at the extreme right end of the histogram.

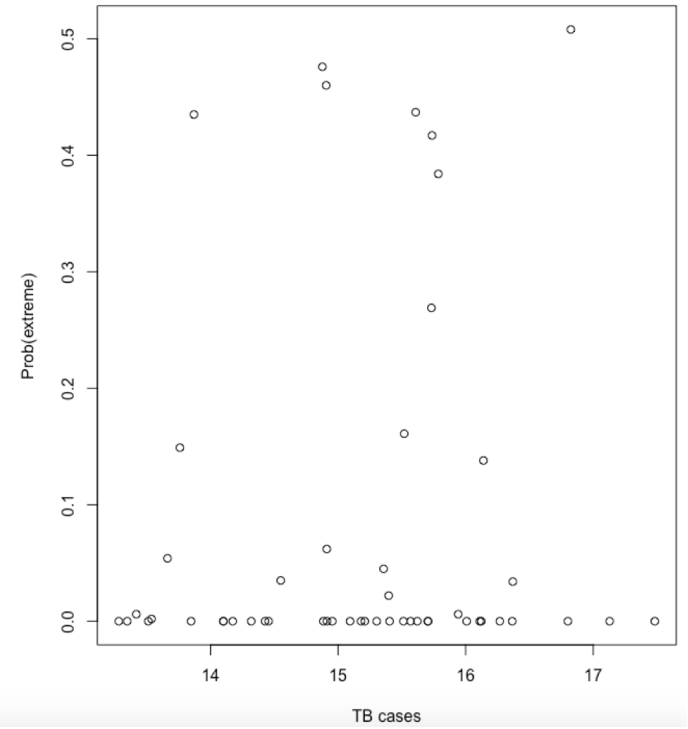

Figure 7: Statewise probability for TB cases

In order to check the consistency of state-wise TB (x_i) cases with its predictive distribution we compute and plot the probability that the future observed x_i^* at least as extreme as x_i by (6) in figure (7). It appears that most of these tail probabilities (40 states are smaller than 0.10) are smaller.

The above three situations indicate the poor fit of the model with equal TB rates and we should consider differences among the rates, and which will be modeled exchangeable model described by (7)-(13).

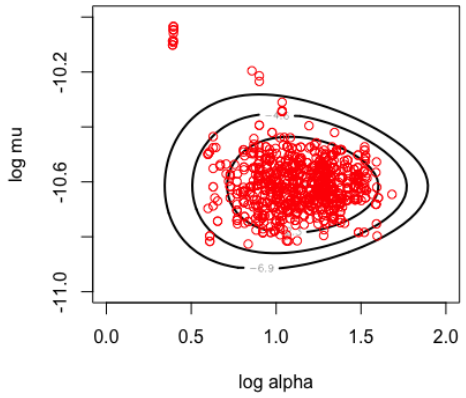

Figure 8: Contour plot of posterior density of $(log\alpha, log\mu)$

The contour plot of posterior density of $(log\alpha, log\mu)$ is given in figure (8). Since most of the points fall within the first two contour lines of the graph, it indicates that the algorithm is giving representative sample form the posterior marginal distribution of $(log\alpha, log\mu)$. First contour line includes 90% of the observations and the second contour line contains 99% observations.

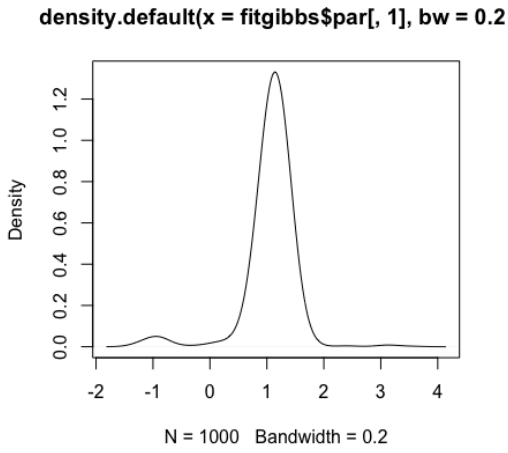

Figure 9: Kernel density estimate of the simulated draws from marginal distribution of $(log\alpha.)$

Figure (9) demonstrates the kernel density estimate of the simulated draws from marginal distribution of $(log\alpha.)$ which is approximately normal with heavy tail at the both end. However; both the contour plot and kernel indicate a better fit of the marginal posterior distributions of the $(log\alpha, log\mu)$.

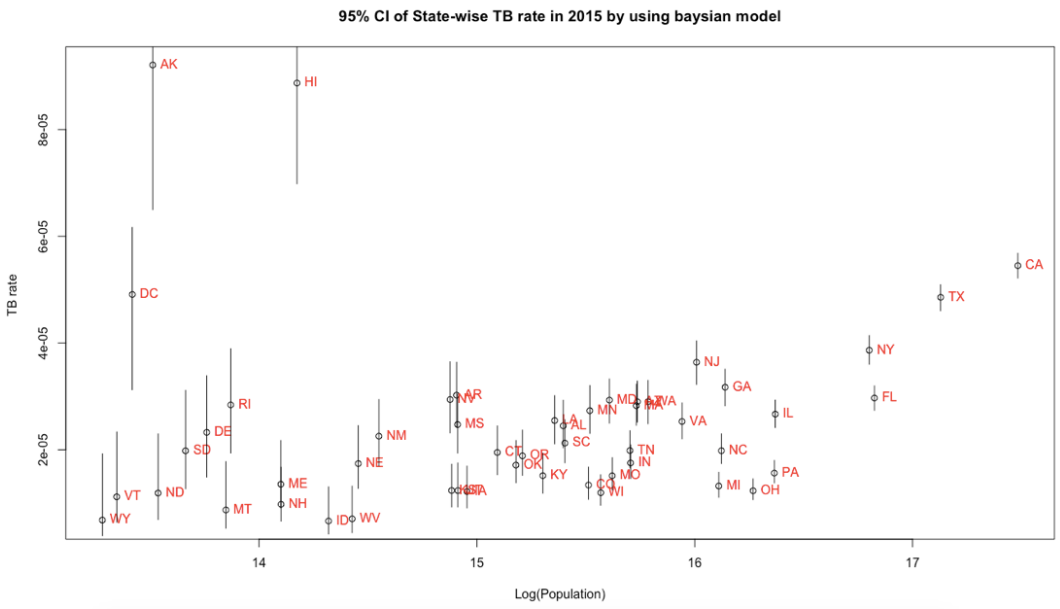

Figure 10: 95% confidence interval of statewise TB rates in 2015

Having the distribution as well as estimates of α and μ one can obtain independent posterior distributions of $\lambda_i s$ as $\gamma(x_i + \alpha, z_i + \frac{\alpha}{\mu})$. With the given α and μ we simulate 1000 estimates of each λ_i and then compute length of middle 95% of estimates and graph them as vertical lines in figure(10). This figure shows that the low observed rates move up and the high observed rates move down which means that the Bayesian estimates of state wise TB rates shrinkage towards pooled rate.

ID	State	z	x	state_abbr	rate	better
46	Vermont	626042	7	VT	1.118136e-05	0.119
47	Virginia	8382993	212	VA	2.528930e-05	0.000
48	Washington	7170351	208	WA	2.900834e-05	0.000
49	West Virginia	1844128	13	WV	7.049402e-06	0.474
50	Wisconsin	5771337	69	WI	1.195564e-05	0.063
51	Wyoming	586107	4	WY	6.824692e-06	0.341

Table 1: Partial view of states TB rates comaprison

We compare the state wise TB rate with the best one (lowest rate state Idaho) and determine the probabilities of being better than Idaho. We observe that the chance of TB rate being better in Wyoming and West Virginia than Idaho is 34.1% and 45.4% respectively.

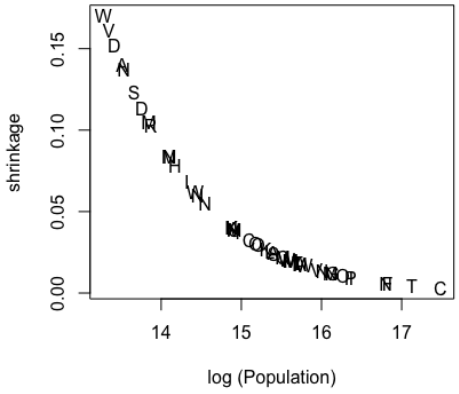

Figure 11: Posterior shringkages for TB cases

Figure(11) depicts that observed rate with low population states shrinkage about 10-20% and states with high population shrinkage around 5% towards pooled estimates.

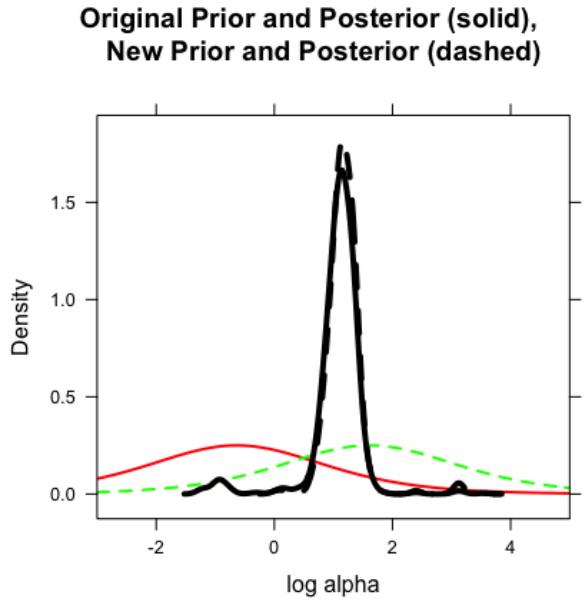

Figure 12: Prior and posterior distribution of

Figure (12) illustrates that with the original prior with $Z_0 = 0.53$ the posterior is the solid line and with new prior with $Z_0 = 5$, the posterior is the dashed line. Both of the posteriors almost identical which indicate that our exchangeable Bayesian model is not sensitive to the change of priors.

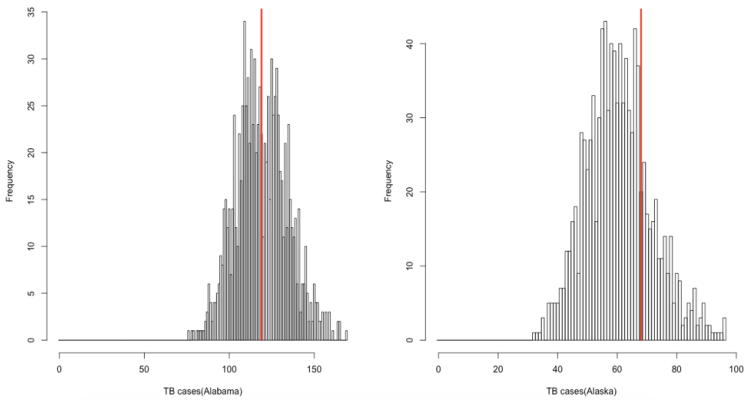

Figure 13: Posterior predictive distribution of future number of TB cases

Figure 13, illustrates that posterior predictive distribution of future number of TB cases in the state Alabama and Alaska fit better in the current model setting exchangeable in comparing to the previous model setting with equal prior.

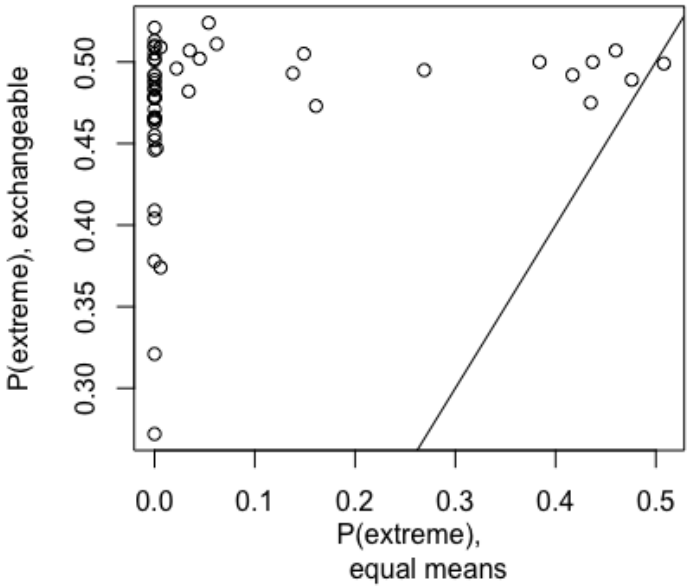

Figure 14: Posterior predictive probabilities of "at least as extreme" for the equal means and exchangeable models

In figure 14, we check the consistency of by constructing scatter plot of posterior predictive probabilities of "at least as extreme" for the equal means and exchangeable models. We notice that probabilities for the exchangeable model are larger than equal mean one indicating that the observations are consistent with the exchangeable fitted model.

Conclusion

This study endeavors to model Tuberculosis (TB) prevalence in the United States across 51 states. In particular, estimating the true TB rates across the estates. The simplest estimates are the ratios of number of TB cases to population sizes of the respective states. However; the estimates might be under estimated or overestimated in case of the states with lower population size. On the other hand, pooled estimate might not be a good choice because of perceived variation in cases and population across states. Another alternative like combining both individual and pooled estimates would improve the estimates. Which is essentially a shrinkage estimate. Bayesian approach might be helpful in this situation. We assume Poisson TB cases and non-informative equal rate prior. Latter on we consider Bayesian exchangeable model by considering the prior of rates as gamma distribution. We estimated the 95% CI of each of the rates and their corresponding shrinks. We observed that rates for the states with lower population size shrink more ($\geq$ 10%) than the states with high population sizes (<5%) towards pooled estimates.

Model performance checking with predictive posterior distribution of future TB cases, contour plot, extreme probability plotting, and consistency checking clearly indicate that Bayesian exchangeable model produces stable results. The data we used for this analysis might not be suitable for this kind of model fitting. Sample data might be appropriate one for this setting where estimating true TB rates would be the main objectives. However, this study intended to apply Bayesian model fitting in practical situations where learning was the priority as compare to information extraction.

Acknowledgement

Kerrigan Research Minigrant

Referance

[**1**] https://www.tbfacts.org/tb-statistics-united-states/
[**2**] Albert, Jim. Bayesian computation with R. Springer Science & Business Media, 2009

Appendix

```
library(dplyr)
mydata<-read.csv("newdata1.csv")
head(mydata)
mydata$e<-as.numeric(gsub('\\,','',mydata$Population))
mydata$y<-as.numeric(gsub('\\,','',mydata$Cases.2015))
attach(mydata)
head(mydata)

mydat2<-subset(mydata,select=c(ID,State,e,y_15,state_abbr))
names(mydat2)<-c("ID","State",'z','x','state_abbr')
attach(mydat2)
summary(log(mydat2$z))
# Individual
plot(log(mydat2$z), (mydat2$x/mydat2$z),  xlab="Log(Population)",
     ylab="TB rate ",
          main='Statewise TB rate comparison in 2015',pch=20,col='red')
text(log(mydat2$z),(mydat2$x/mydat2$z),
     labels=mydat2$state_abbr,pos=4,col='black',cex=1)
# Combined
sum_z<-sum(mydat2$z)
sum_x<-sum(mydat2$x)
combine_rate<-(sum_x/sum_z)
combine_rate
# Equla mortality rate
lambda=rgamma(1000,shape=sum_x,rate=sum_z)
# for the first state Alabama
set.seed(1)
ys1=rpois(1000,mydat2$z[1]*lambda)
mean(ys1)
hist(ys1,xlab='TB cases',main='Posterior predictive distribution of TB cases in
     Alabama ',breaks=seq(0.5,max(ys1)+0.5),)
lines(c(y[1],y[1]),c(0,120),lwd=3)
# for the second state Alaska
set.seed(1)
ys2=rpois(1000,mydat2$z[2]*lambda)
mean(ys2)
hist(ys2,breaks=seq(1,70,by=2))
lines(c(y[2],y[2]),c(0,120),lwd=3)
set.seed(1)
lambda=rgamma(1000,shape=sum_x,rate=sum_z)
lambda
prob.out=function(i)
  {
     ysi=rpois(1000,mydat2$z[i]*lambda)
    pleft=sum(ysi<=mydat2$x[i])/1000
     pright=sum(ysi>=mydat2$x[i])/1000
    min(pleft,pright)
}


pout=sapply(1:51,prob.out)
```

```
pout
plot(log(mydat2$z),pout,ylab="Prob(extreme)",xlab='TB cases')
datapar = list(data =mydat2, z0 = 0.53)
start=c(2, -7)
poissgamexch=function (theta, datapar)
{
  x= datapar$data[,4]
  z = datapar$data[,3]
  z0 = datapar$z0
  alpha = exp(theta[1])
  mu = exp(theta[2])
  beta = alpha/mu
  logf = function(x, z, alpha, beta)
    lgamma(alpha + x) - (x + alpha) * log(z + beta) +
    alpha * log(beta) - lgamma(alpha)
  val = sum(logf(x, z, alpha, beta))
  val = val + log(alpha) - 2 * log(alpha + z0)
  return(val)
}
library(LearnBayes)
fit = laplace(poissgamexch, start, datapar)
par(mfrow = c(1, 1))
mycontour(poissgamexch, c(0,2,-11,-10), datapar,xlab="log alpha",ylab="log mu")
start = c(4, -7)
fitgibbs = gibbs(poissgamexch, start, 1000, c(1,.15), datapar)
fitgibbs$accept
mycontour(poissgamexch, c(0,2,-11,-10), datapar,
          xlab="log alpha",ylab="log mu")
points(fitgibbs$par[, 1], fitgibbs$par[, 2],col='red')
plot(density(fitgibbs$par[, 1], bw = 0.2))
alpha = exp(fitgibbs$par[, 1])
mu = exp(fitgibbs$par[, 2])
lam1 = rgamma(1000, mydat2$x[1] + alpha, mydat2$z[1] + alpha/mu)
alpha = exp(fitgibbs$par[,1])
 mu = exp(fitgibbs$par[,2])
plot(log(mydat2$z), (mydat2$x/mydat2$z), pch =as.character(state_abbr))
plot(log(mydat2$z), (mydat2$x/mydat2$z),  xlab="Log(Population)",
     ylab="TB rate ",
     main='95% CI of State-wise TB rate in 2015 by using baysian model')
text(log(mydat2$z),(mydat2$x/mydat2$z),
     labels=mydat2$state_abbr,pos=4,col='red')
for (i in 1:51) {
   lami = rgamma(1000, mydat2$x[i] + alpha, mydat2$z[i] + alpha/mu)
   probint = quantile(lami, c(0.025, 0.975))
   lines(log(mydat2$z[i]) * c(1, 1), probint)
   }
# Srinkage
shrink=function(i) mean(alpha/(alpha +mydat2$z[i] * mu))
shrinkage=sapply(1:51, shrink)
plot(log(mydat2$z), shrinkage,xlab='log (Population)', pch =as.character(mydat2$State))
Tbrate=function(i) mean(rgamma(1000, mydat2$x[i] + alpha, mydat2$z[i]+alpha/mu))
state=1:51
meanrate=sapply(state,Tbrate)
```

```
state[meanrate==min(meanrate)]
mydat2$rate<-(mydat2$x/mydat2$z)
order(mydat2$rate)
#comparing state
sim.lambda=function(i) rgamma(1000,mydat2$x[i]+alpha,mydat2$z[i]+alpha/mu)
LAM=sapply(1:51,sim.lambda)
compare.rates <- function(x) {
  nc <- NCOL(x)
  ij <- as.matrix(expand.grid(1:nc, 1:nc))
  m <- as.matrix(x[,ij[,1]] > x[,ij[,2]])
  matrix(colMeans(m), nc, nc, byrow = TRUE)
  }
better=compare.rates(LAM)
dim(better)
mydat2$better<-better[1:51,13]
mydat2

#Baysian sensitivity analysis
sir.old.new=function(theta, prior, prior.new)
{
  log.g=log(prior(theta))
  log.g.new=log(prior.new(theta))
  wt=exp(log.g.new-log.g-max(log.g.new-log.g))
  probs=wt/sum(wt)
  n=length(probs)
  indices=sample(1:n,size=n,prob=probs,replace=TRUE)
  theta[indices]
}

prior=function(theta)
  0.53*exp(theta)/(exp(theta)+0.53)^2
prior.new=function(theta)
  5*exp(theta)/(exp(theta)+5)^2

log.alpha=fitgibbs$par[, 1]
log.alpha.new=sir.old.new(log.alpha, prior, prior.new)

############ drawing figure

library(lattice)
draw.graph=function()
{
  LOG.ALPHA=data.frame("prior",log.alpha)
  names(LOG.ALPHA)=c("Prior","log.alpha")
  LOG.ALPHA.NEW=data.frame("new.prior",log.alpha.new)
  names(LOG.ALPHA.NEW)=c("Prior","log.alpha")
  D=densityplot(~log.alpha,group=Prior,data=rbind(LOG.ALPHA,LOG.ALPHA.NEW),
                plot.points=FALSE,main="Original Prior and Posterior (solid),
   New Prior and Posterior (dashed)",
                lwd=4,adjust=2,lty=c(1,2),xlab="log alpha",xlim=c(-3,5),col="black")
  update(D, panel=function(...){
    panel.curve(prior(x),lty=1,lwd=2,col="red")
    panel.curve(prior.new(x),lty=2, lwd=2,col="green")
```

```
    panel.densityplot(...)
  })}
draw.graph()
#posterior predictive model checking
par(mfrow=c(1,2))
lam51=rgamma(1000,mydat2$x[1]+alpha,mydat2$z[1]+alpha/mu)
ys51=rpois(1000,mydat2$z[1]*lam51)
hist(ys51,xlab='TB cases(Alabama)',main='',breaks=seq(-0.5,max(ys51)+0.5))
lines(mydat2$x[1]*c(1,1),c(0,100),lwd=3,col='red')
lam51=rgamma(1000,mydat2$x[2]+alpha,mydat2$z[2]+alpha/mu)
ys51=rpois(1000,mydat2$z[2]*lam51)
hist(ys51,xlab='TB cases(Alaska)',main='',breaks=seq(-0.5,max(ys51)+0.5))
lines(mydat2$x[2]*c(1,1),c(0,100),lwd=3,col='red')
prob.out=function(i)
  {
     lami=rgamma(1000,mydat2$x[i]+alpha,mydat2$z[i]+alpha/mu)
    ysi=rpois(1000,mydat2$z[i]*lami)
    pleft=sum(ysi<=mydat2$x[i])/1000
    pright=sum(ysi>=mydat2$x[i])/1000
    min(pleft,pright)
    }
pout.exchange=sapply(1:51,prob.out)
plot(pout,pout.exchange,xlab="P(extreme),
     equal means",ylab="P(extreme), exchangeable")
abline(0,1)
tail(mydat2)
```