Heirarchical Modelling on co-morbidity counts in smoking and non-smoking heart failure patients

Alden Tan

22nd March 2022

Introduction

Heart failure is the end-stage condition that occurs when the heart muscle no longer has sufficient strength to pump blood around the body, which may be due to mechanical weakness or other underlying co-morbidities, such as coronary artery occlusion, high blood pressure or anatomical defects. This condition affects millions worldwide as prevalence of cardiovascular diseases increase due to a global change in lifestyles and consumption habits. Co-morbidities are auxiliary conditions such as high blood glucose, high blood pressure, obesity and so on, and these conditions contribute to heart failure outcomes in patients with cardiovascular disease. Another important factor, smoking status, also plays a role in determining heart failure of patients with cardiovascular disease, but is a lifestyle choice rather than a co-morbidity. Smoking has long been documented to affect many aspects of a patient's physiology, like blood glucose and pressure, among other things. Strong associations between smoking and co-morbidities are established as well. Thus, in order to better understand the health profile of patients with heart disease, we will examine the number of co-morbidities in 2 groups of individuals that experience heart failure, one group consisting of smokers while the other group consists of non-smokers.

Source Material and Data

The data used in this analysis was obtained from the UCI Machine Learning Repository, titled as "Heart failure clinical records Data Set". It contains information about 299 individuals with class III or class IV heart failure recorded at the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad. Thirteen points of clinical information was collected, with our main interest being: smoking, anemia, blood pressure and diabetes status (binary), blood platelet level (platelets/ul) and serum creatinine level (mg/dL). Variables of binary status were unchanged, while blood platelet measurement was categorized into abnormal/normal platelet count (1/0) using the normal range of 150,000 to 450,000 platelets/ul. Serum creatinine level was also categorized in to abnormal/normal kidney (1/0) status based on the normal range of 0.84 to 1.21mg/dL. Individuals were categorized by smoking status and the five binary variables (anemia, blood pressure, kidney, platelet and diabetes status) were summed up for each individual.

Exploring our Data

```
## Data Description
library(coda)
cover<-read.table("C:/Users/tanal/Downloads/smoking_comorbidity.txt",header=T
)
Y<-cover$All_comorbidities
old_data<-density(Y)

hist(Y,freq=F,breaks=4,main ="Histogram of comorbidity data", xlab = "Number
of co-morbidities",ylim=c(0,1))
lines(x=old_data$x,y=old_data$y,col=2)
```

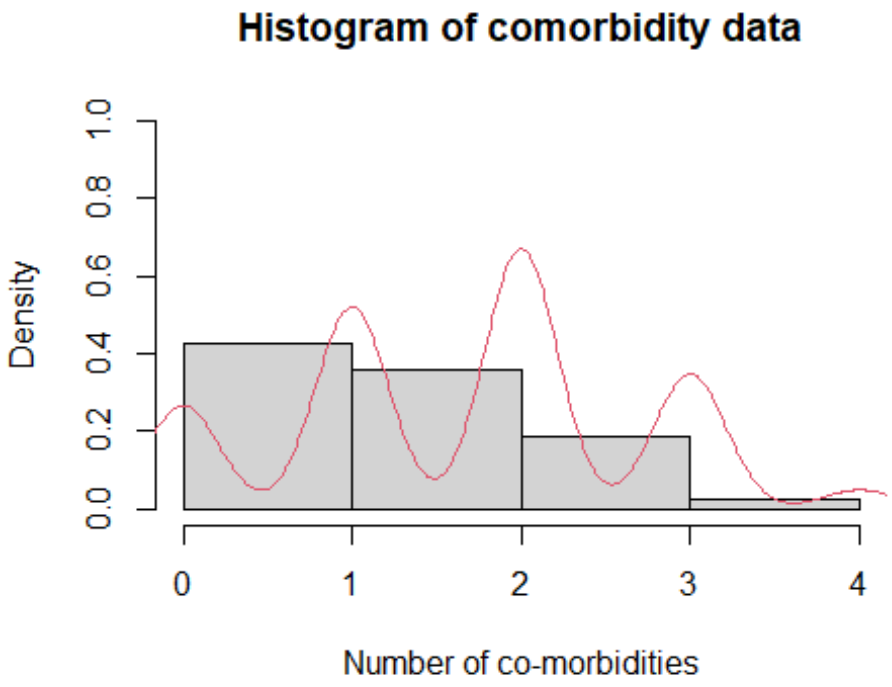

Figure 1: Red line indicates kernel density of co-morbidities counts for 299 individuals in the dataset.

Data Summary Statistics

```
mean(Y) # Overall mean

## [1] 1.672241

var(Y) # Data variance

## [1] 1.046576
```

Deriving Posterior Distribution

Given:

$$Y_{ij} \sim Poisson(\lambda_j \mu)$$

$$\lambda_j \sim Gamma(c, c)$$

$$\mu \sim Gamma(a, b)$$

PDF of y:

$$p(y_{ij}|\mu, \boldsymbol{\lambda}) \propto \frac{(\lambda_j \mu)^y e^{-\lambda_j \mu}}{y_{ij}!}$$

Likelihood of data, $\mathbf{y}$:

$$p(\mathbf{y}|\mu, \boldsymbol{\lambda}) \propto \Pi_{j=1}^{m} \Pi_{i=1}^{n_j} \frac{(\lambda_j \mu)^{yij} e^{-\lambda_j \mu}}{y_{ij}!}$$

Priors:

$$p(\mu) \propto \mu^{a-1} e^{-b\mu}$$

$$p(\lambda_j) \propto \lambda_j^{c-1} e^{-c\lambda_j}$$

Since each λ_j, μ is independent:

$$p(\mu, \boldsymbol{\lambda}) \propto \mu^{a-1} e^{-b\mu} \Pi_{j=1}^{m} \lambda_j^c e^{-c\lambda_j}$$

Posterior Distribution:

$$p(\boldsymbol{\lambda}, \mu|\mathbf{y}) = p(\mathbf{y}|\mu, \boldsymbol{\lambda}) \cdot p(\mu, \boldsymbol{\lambda})$$

$$\propto \Pi_{j=1}^{m} \Pi_{i=1}^{n_j} \frac{(\lambda_j \mu)^{yij} e^{-\lambda_j \mu}}{y_{ij}!} \cdot \Pi_{j=1}^{m} \lambda_j^c e^{-c\lambda_j} \cdot \mu^{a-1} e^{-b\mu}$$

Posterior of μ:

$$p(\mu|\boldsymbol{\lambda}, \mathbf{y}) \propto \Pi_{j=1}^{m} \Pi_{i=1}^{n_j} \mu^{y_{ij}} e^{-\lambda_j \mu} \cdot \mu^{a-1} e^{-b\mu}$$

$$= \mu^{a + \sum_{j=1}^{m} \sum_{i=1}^{n_j} y_{ij} - 1} \cdot exp\{-\mu(b + \sum_{j=1}^{m} n_j \lambda_j\}$$

$$p(\mu|\boldsymbol{\lambda}, \mathbf{y}) \sim Gamma\left(a + \sum_{j=1}^{m} \sum_{i=1}^{n_j} y_{ij}, b + \sum_{j=1}^{m} n_j \lambda_j\right)$$

Posterior of λ_j:

$$p(\lambda_j|\mu,\mathbf{y}) \propto \lambda_j^{c-1} e^{-c\lambda_j} \Pi_{i=1}^{n_j} (\lambda_j)^{y_{ij}} e^{-\lambda_j \mu}$$

$$\propto \lambda_j^{c-1+\sum_{i=1}^{n_{ij}} y_{ij}} \cdot exp\{-\lambda_j \left(c_{n_j}\mu\right)\}$$

$$p(\lambda_j|\mu,\mathbf{y}) \sim Gamma\left(c + \sum_{i=1}^{n_j} y_{ij}\,, c + n_j\mu\right)$$

Performing Gibbs Sampling

```
## Performing Gibbs Sampling
mu_0<-mean(cover$All_comorbidities) # overall mean
Y<-cover$All_comorbidities # Data
n_total<-length(cover$smoking) # Total data size
n1=sum(cover$smoking==0) # number of non-smoking
n2=sum(cover$smoking==1) # number of smoking

non_smoking<-cover$All_comorbidities[cover$smoking==0]
smoking<-cover$All_comorbidities[cover$smoking==1]
scaling_factor_smoking<-mean(smoking)/mu_0 #Scaling mu_0 to smoking group mea
n
scaling_factor_non_smoking<-mean(non_smoking)/mu_0 #Scaling mu_0 to non-smoki
ng group mean

a=mu_0
b=1
c<-mean(c(scaling_factor_smoking,scaling_factor_non_smoking)) ### Middle valu
e of the two means
K=10000

Parameter<-matrix(0, nrow=K,ncol = 3)

for (i in 1:K){

  lambda_1_n<-rgamma(1,c+sum(non_smoking), c+ (n1*mu_0))
  lambda_2_n<-rgamma(1,c+sum(smoking), c+ (n2*mu_0))
  u_n<-rgamma(1,a+sum(Y),b+sum(n1*lambda_1_n,n2*lambda_2_n))

  Parameter[i,]<-c(u_n,lambda_1_n,lambda_2_n)

  mu_0<-u_n ##Update
}
u_post<-Parameter[,1]
lambda_post_non_smoking<-Parameter[,2]
lambda_post_smoking<-Parameter[,3]
```

```r
breakpoint_1<-n1/n_total
new_Y<-c()
new_smoker<-c()
new_non_smoker<-c()
for (i in 1:K){
  x_value<-runif(1,0,1)
  if (x_value < breakpoint_1){
    new_Y<-c(new_Y,rpois(1,lambda_post_non_smoking[i]*u_post[i]))
    new_non_smoker<-c(new_non_smoker,new_Y[i])
  }
  else{
    new_Y<-c(new_Y,rpois(1,lambda_post_smoking[i]*u_post[i]))
    new_smoker<-c(new_smoker,new_Y[i])
  }
}

new_Y_density<-density(new_Y)
hist(x=new_Y,freq = F,main="Sample from Posterior Predictive Distribution",xl
ab="Number of co-morbidities",ylim=c(0,1.5))
lines(x=new_Y_density$x,y=new_Y_density$y,col=4)
lines(x=old_data$x,y=old_data$y,col=2)
```

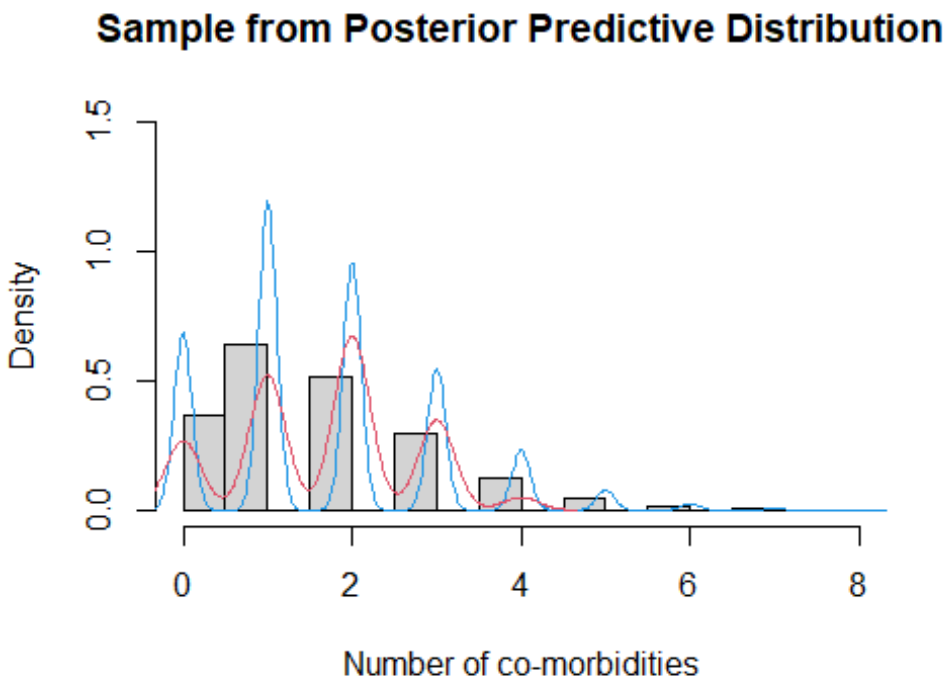

Figure 2: Red line indicates kernel density of the co-morbidity count of 299 individuals in the dataset. Blue line indicates kernel density of the number of co-morbidities after sampling 10,000 times from the posterior predictive distribution.

```
plot(mcmc(u_post*lambda_post_non_smoking))
```

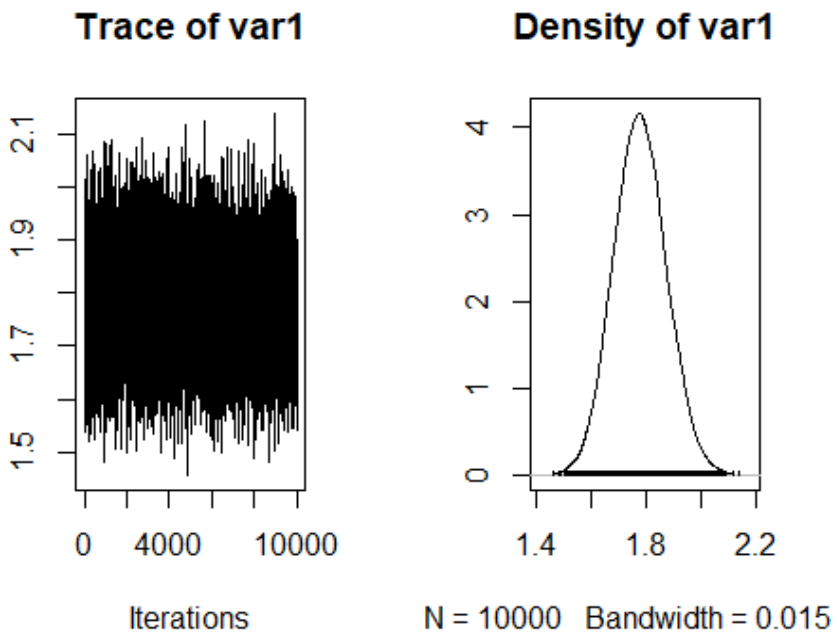

Figure 3: MCMC diagnostic plot for posterior distribution of μ*λ for non-smoking group.

```
plot(mcmc(u_post*lambda_post_smoking))
```

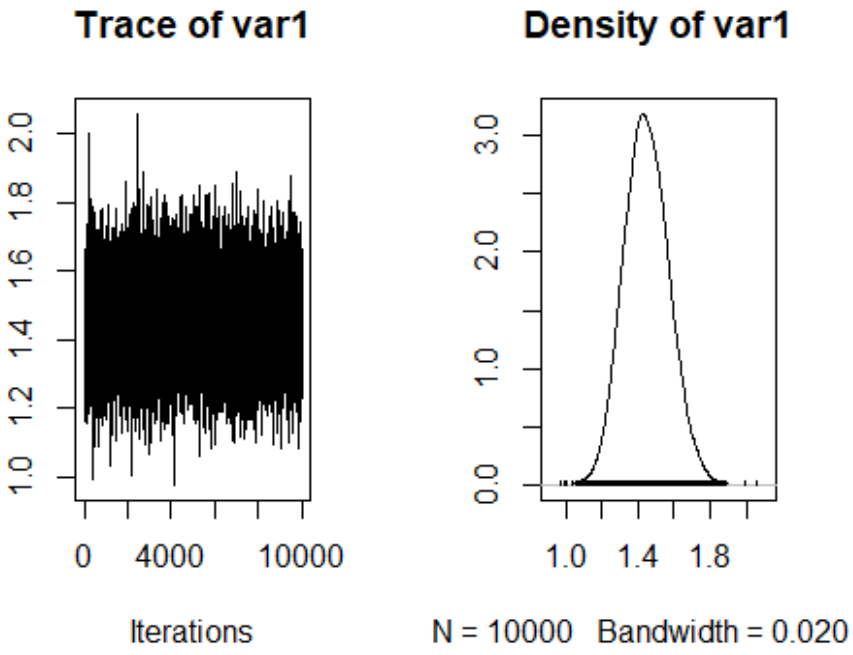

Figure 3: MCMC diagnostic plot for posterior distribution of μ*λ for smoking group.

Parameter(s) Summary Statistics

```
## Posterior Means
mean(u_post) # u_n

## [1] 1.927485

mean(lambda_post_non_smoking) # lambda_non-smoking

## [1] 1.136376

mean(lambda_post_smoking) #lambda_smoking

## [1] 0.9248302

## 95% Credible Interval
quantile(u_post,c(0.025,0.975))

##      2.5%     97.5%
## 0.7042193 4.3088307

quantile(lambda_post_non_smoking,c(0.025,0.975))

##      2.5%     97.5%
## 0.4144128 2.5261436

quantile(lambda_post_smoking,c(0.025,0.975))

##      2.5%     97.5%
## 0.3331719 2.0309884

## Group mean
mean(new_non_smoker)

## [1] 1.751497

mean(new_smoker)

## [1] 1.493177
```

Results

The results of hierarchical modelling are quite intriguing. First, from figure 2, we can see that setting hyperparameters a= 1.672, b=1 and c= 0.964 did not follow the density trend plotted out in the posterior predictive density, where the original data exhibit minor left skew, the posterior predictive density exhibits right skew. The posterior predictive density is also greater in scale than the original plot for the data, which suggests that selected hyperparameters for a, b and c might not be appropriate for this model.