Denis Ostroushko - PUBH 7440 - HW3

Problem 1

I am attaching a derivation of the conditional posterior distribution for $lambda_{i\alpha}$ at the end of the document as a hand-written part. I am showing how I obtain a posterior gamma distribution based on the distribution of $Y_{i\alpha}$ and prior distribution of $\lambda_{i\alpha}$. Resulting distribution is $Gamma(Y_{0\alpha} + Y_{i\alpha}, n_{0\alpha} + n_{i\alpha})$.

Problem 2

In this version of the Gamma distribution parametrization, the mean is given by $\frac{\alpha}{\beta}$, or $\frac{Y_{0\alpha}}{n_{0\alpha}}$, which is the death rate we want to analyze. Therefore, the whole estimated Gamma distribution is 'centered' at the estimate death rate, and the distribution provides expected variation around the point estimate $\frac{Y_{0\alpha}}{n_{0\alpha}}$.

Problem 3

This makes sense because we will preserve the distribution of age groups within a county which was observed in the data. Death rate will be based on a proportional population for age group α, according to values of parameter $\lambda_{0\alpha}$. This way we can control the value of total country population, and though $\pi_{0\alpha}$ we control the number of people in each age group.

Problem 4

In order to impute missing/suppresed values of $Y_{i\alpha}$ we need to use a truncated left tail of a poisson distribution with corresponding parameter $n_{i\alpha}\lambda_{i\alpha}$. We will set a maximum value at the tail equal to 10, meaning that for our imputations we will be sampling integers from 0 to 10 from poisson distributions. In order to do that, we follow these steps:

1. For each county for each group age, determing a parameter for the poisson distribution, refer to it as $\Lambda_{i\alpha}$.
2. For each county for each age group, determine quantile corresponding to value of 10 under $\Lambda_{i\alpha}$, call this quantile q

- use `ppois()` to get this quantile

3. Sample a number from a uniform distribution between 0 and q. This will be between 0 and some number less than or equal to 1 always.

- use `runif(n=1, min = 0, max = .)`

4. Using inverse CDF of a poisson distribution with parameter $\Lambda_{i\alpha}$, obtain a value correspoding to a randomly sampled quantile

- use `qpois()` for this step

5. Impute missing value with sampled values between 0 and 10.
6. Using imputed data, obtain posterior estiamtes on the number of death and population size and sample new rates from $Gamma(Y_{0\alpha} + Y_{i\alpha}, n_{0\alpha} + n_{i\alpha})$.

Problem 5

We want to learn about the death rates in each county in each age group. Recall that $\lambda_{i\alpha}$ represents mortality rate associated with stroke in each county $i = 1, 2, \ldots, 67$ in each age group $\alpha = 1, 2, 3$. In the Bayesian data analysis framework, we want to obtain a posterior distribution of each parameter $\lambda_{i\alpha}$ given observed death rates, or death counts (the data) $Y_{i\alpha}$ and population size corresponding to an age group in the county i.

In the framework of our analysis, we treat population size for age group α in county i as a constant value.

According to the *Problem 1* statement, the likelihood for observed data is $Y_{i\alpha}$~$Pois(n_{i\alpha}\lambda_{i\alpha})$, and the prior distribution of the parameter of interest is $\lambda_{i\alpha}$~$Gamma(n_{0\alpha}, Y_{0\alpha})$.

Additionally, because of the suppressed data, we need to specify likelihood of these censored Y values.

So, $p(\lambda_{i\alpha}|Y_{i\alpha}, n_{i\alpha}, Y_{0\alpha}, n_{0\alpha}) \propto \Pi_{observed\ death} Pois(n_{i\alpha}\lambda_{i\alpha}) \times \Pi_{suppresed\ death} F(10|n_{i\alpha}\lambda_{i\alpha}) \times Gamma(Y_{0\alpha}, n_{0\alpha})$

Problem 6

Gibbs sampler outline:

1. Initiate $\lambda_{i\alpha}$ at 75, 250, 1000 deaths per 100,000 for each age group respectively
2. Set prior guess at the population size at each age group within each county with total $n_0 = 10,000$ and corresponding π_α
3. Set prior guess at the death number for each age group using using prior population size and prior death rate
4. Begin Gibbs Sampling. I am using 10,000 iterations.
5. Impute the data:

- at iteration 1, impute data using process described in *Problem 4* and prior guesses of $\lambda_{i\alpha}$
- at iterations 2, 3, ...,10,000 use most recent sampled value of $\lambda_{i\alpha}$

6. Using imputed (complete) data get parameters for posterior distribution of $\lambda_{i\alpha}$ and sample new values for the next iteration of gibbs sampling

Code for execution of the sampler is given below. I wrote my own version of R code for this taks:

```
reps = 10000

results <- cbind(matrix(data = NA,
                        nrow = nrow(stroke_clean),
                        ncol = reps),
                 stroke_clean %>% select(county, age.group)
                 ) # empty matrix for results

results[,1] <- stroke_clean$lambda_0 # itiate sampler with prior guesses of lambdas

set.seed(178921)

for(i in 2:reps){

  if(i %% 1000 == 0){print(i)}
  # impute missing values of Y using inverse CDF approach
    # use previous estimates of lambda parameters to get rate for the poisson distribution
    results[,(i-1)] * stroke_clean$population -> poisson_lambdas_iter

    ppois(10, poisson_lambdas_iter) -> limits_detection_iter
```

```r
    # using these numbers between 0 and somewhere less than 1, sample from uniform distrib
    runif(n = length(limits_detection_iter), min = 0, max = limits_detection_iter) -> samp

    # get imputed values by putting unifrom random samples into 'inverse' CDF
    qpois(sampled_u, lambda = poisson_lambdas_iter) -> imp

    # get final imputed vector of the observed data
    stroke_clean$deaths -> final_ys_iter
    final_ys_iter[which(is.na(final_ys_iter))] <- imp[which(is.na(final_ys_iter))]

  # now work with prior n0 Y0 and observed n_ia Y_ia to get samples for parameters lambda

    pop = stroke_clean$population

    rgamma(n = nrow(stroke_clean),
    #       shape = final_ys_iter + results[,(i-1)]*pop,  # old version
           shape = final_ys_iter + with(stroke_clean, lambda_0 * n_0),
           scale = 1/with(stroke_clean, population + n_0)
           ) -> results[,i]
}

write_rds(results, "gibbs results.rds")
```

Problem 7

Figure 1 is the resulting map

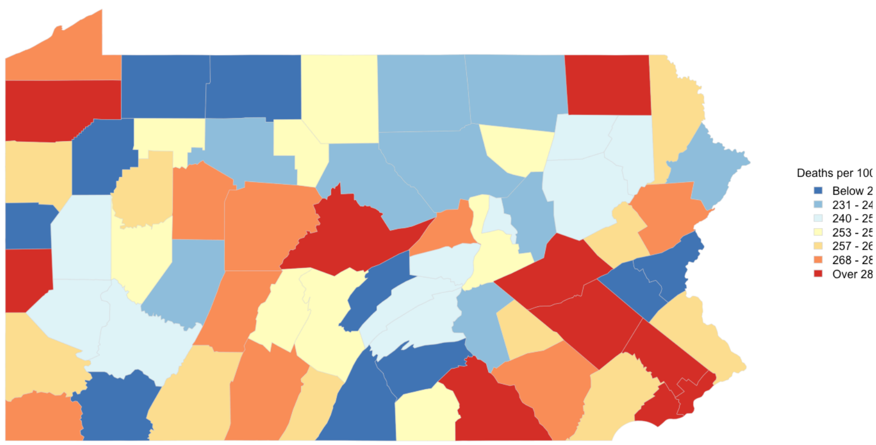

Figure 1: Final Map of Rates

Problem 1

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

- $i \Rightarrow$ county number

  $\alpha \Rightarrow$ age group.

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

  $n_{i\alpha}$ = population

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

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

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

- Posterior:

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

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

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

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

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

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

So, a full conditional distribution can be written as

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

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