Study 3. Nongenetic Birth Defect Disease Rate - Poisson - MH

Ting Lu

2023-10-24

Introduction

This study investigating the impact of PCB Contamination to disease rate in six neighboring counties over a five-year period. **The number of occurrences of a rare, nongenetic birth defect** in a five-year period for 6 neighboring counties is y = (1, 3, 2, 12, 1, 1). The counties have populations of 2 = (33, 14, 27, 90, 12, 17), given in thousands. The second county has higher rates of toxic chemicals (PCBs) present in soil samples, and it is of interest to know if this town has a high disease rate as well.

Install Package

```
# library(tidyverse): data manipulation & visualization
# library(MASS): statistical methods & regression analysis
# library(MCMCpack): MCMC simulations &  Bayesian statistical modeling and parameter estimation
```

`Tidyverse` includes multiple packages: `ggplot2`, `dplyr`, `tidyr`, `readr`, `purrr`, `tibble`, `stringr`

Hierarchical Poisson Model & Prior setting

$$
\begin{aligned}
&Y_i|\theta_i, x_i \sim \text{Poisson}(\theta_i x_i);\\
&\theta_1, \ldots, \theta_6|a, b \sim \text{gamma}(a, b);\\
&a \sim \text{gamma(1,1)} \ ; \ b \sim \text{gamma(10,1)}.
\end{aligned}
$$

- Y_i: the number of occurrences of disease for 6 neighboring counties, therefore, y_i/x_i represents the observed disease rate.

- θ_i: represents the expected disease rate where $E(\theta_i) = a/b$.
- a and b are the parameters of the gamma distribution sampling model of θ_i.

Conditional distribution of $p(\theta_1, ..., \theta_6|a, b, x, y)$:

$$p(\theta_1, ..., \theta_6|a, b, x, y) = \frac{p(\theta_1, ..., \theta_6, y|a, b, x)}{p(y|a, b, x)}$$

- $p(\theta_1, ..., \theta_6|a, b, x, y)$ is the posterior probability distribution of parameters $\theta_1, ..., \theta_6$ given a, b, x, and observed data y. This is the target we want to estimate, i.e., the distribution of parameters given the observed data and certain conditions.
- $p(\theta_1, ..., \theta_6, y|a, b, x)$ is the joint probability distribution of parameters $\theta_1, ..., \theta_6$ and observed data y given a, b, x.
- $p(y|a, b, x)$ is the marginal probability distribution of observed data y given a, b, x. It represents the probability of the observed data, independent of the parameters.

The posterior probability distribution is equal to the joint probability distribution divided by the marginal probability distribution.

Therefore, the posterior densities is:

$$\begin{aligned} p(\theta_1, ..., \theta_6, y|a, b, x) &= p(\theta_1, ..., \theta_6|a, b) * p(y|\theta_1, ..., \theta_6, x) \\ &= \prod_{i=1}^{6} \frac{(\theta_i x_i)^{y_i} exp(-\theta_i x_i)}{y_i!} * \prod_{i=1}^{6} \frac{(b)^a}{\Gamma(a)} \theta_i^{a-1} exp(-b\theta_i) \\ &= f(y) \prod_{i-1}^{6} \theta_i^{a+y_i-1} exp(-(b + x_i)\theta_i) \\ &\propto \prod_{i-1}^{6} \theta_i^{a+y_i-1} exp(-(b + x_i)\theta_i) \end{aligned}$$

Because θ_i are conditional independent given a, b, x, y, full conditional distribution of the disease rate for each county is:

$$p(\theta_i|\theta_{-i}, a, b, x, y) \propto \prod_{i-1}^{6} \theta_i^{a+y_i-1} exp(-(b + x_i)\theta_i) \sim gamma(a + y_i, b + x_i)$$

Acceptance Ratio The ratio of the posterior densities comparing a set of **proposal** values (a^*, b^*, θ) to values (a, b, θ):

$$\begin{aligned} r &= \frac{p(\theta_1, ..., \theta_6, y|a^*, b^*, x)}{p(\theta_1, ..., \theta_6, y|a, b, x)} \\ &= \prod_{i=1}^{6} \frac{\theta_i^{a^*+y_i-1} exp(-(b^* + x_i)\theta_i)}{\theta_i^{a+y_i-1} exp(-(b + x_i)\theta_i)} \\ &= \prod_{i-1}^{6} \theta_i^{a^*-a} exp(-(b^* - b)\theta_i) \end{aligned}$$

- Note the value of θ, the vector of county-specific rates, is unchanged.

Metropolis-Hastings Sampling (MCMC sampling)

Run Metropolis-Hastings Algorithm to generate samples of (a, b, θ) from the posterior, randomly sampling probabilities from uniform distribution to compare with acceptance ratio.

```r
# observed data
x <- c(33, 14, 27, 90, 12, 17)
y <- c(1, 3, 2, 12, 1, 1)
# prior
a <- 1
b <- 10
theta <- rgamma(6, a + y, b + x)
a_delta <- 1
b_delta <- 10
# Define Metropolis-Hastings Algorithm
ratio <- function(a, b, a_new, b_new, theta) {
  prod(theta^(a_new - a) * exp(-(b_new - b) * theta))
}
S <- 10000
set.seed(123)
THETA <- NULL

# Run Metropolis-Hastings Algorithm
for (j in 1:S) {
  if (j%%(S/10) == 0) {
    print(j)
    }
  a_new <- abs(runif(1, a - a_delta, a + a_delta))
  b_new <- abs(runif(1, b - b_delta, b + b_delta))
  r <- ratio(a, b, a_new, b_new, theta) #compute the acceptance ratio
  if (runif(1) < r) {
    # accept
    theta <- rgamma(6, a_new + y, b_new + x)
    THETA <- rbind(THETA, c(a_new, b_new, theta))
    a <- a_new
    b <- b_new
    } else {
      theta <- rgamma(6, a + y, b + x)
      THETA <- rbind(THETA, c(a, b, theta))
    }
}
```

```
## [1] 1000
## [1] 2000
## [1] 3000
## [1] 4000
## [1] 5000
## [1] 6000
## [1] 7000
## [1] 8000
## [1] 9000
## [1] 10000
```

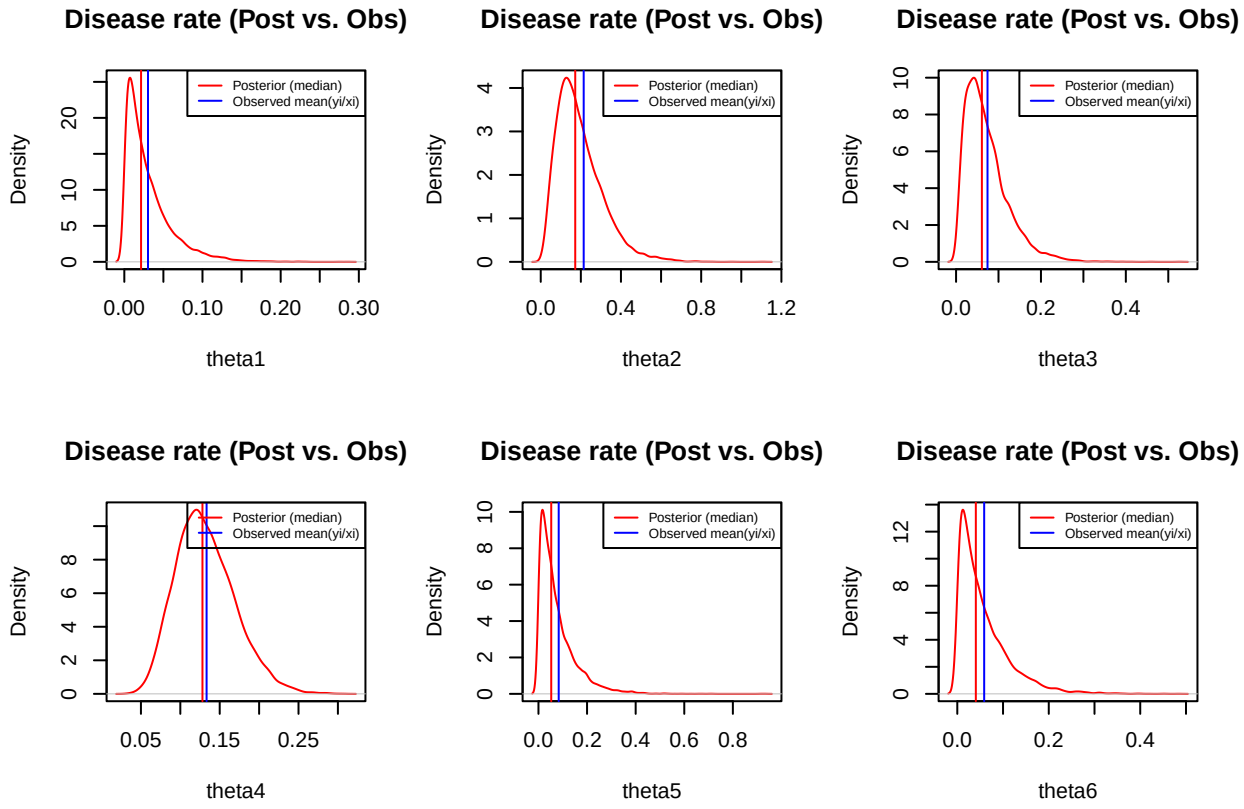
Posterior inference for disease rates

Compare marginal posterior distributions of $\theta_1, ..., \theta_6$ with observed $y_1/x_1, ..., y_6/x_6$

```
burn_in <- 1000
theta.post <- THETA[, -(1:2)]
y.x <- y/x

par(mfrow = c(2,3))
for (j in 1:6) {
  plot(density(theta.post[,j]), xlab = paste0("theta", j),main = paste0("Disease rate (Post vs. Obs)"),
       col='red')
  abline(v = y.x[j],col = 'blue')
  abline(v = median(theta.post[,j]), col = "red")
  legend('topright', legend=c("Posterior (median)", "Observed mean(yi/xi)"),
       col=c("red", "blue"), lty=1,cex=0.6)

}
```



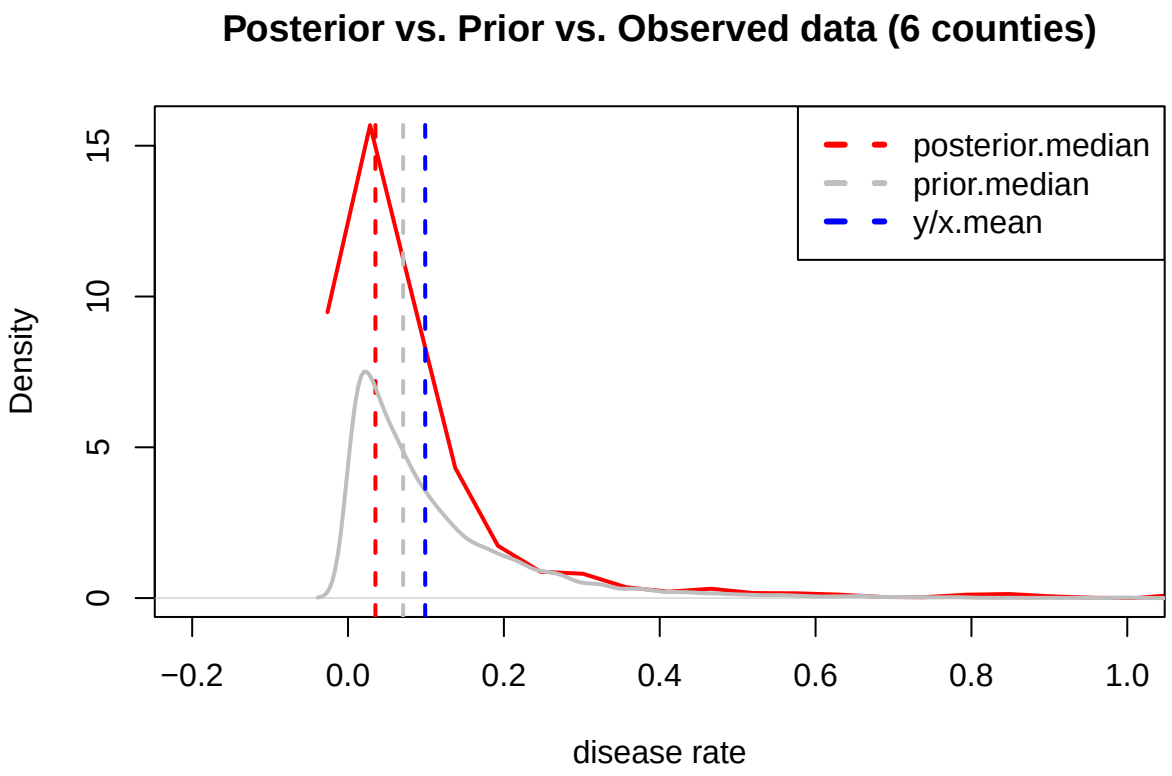
Based on the plots, each observed average disease rate y_i/x_i is larger than the median value of posterior distribution of $\theta_1, ...\theta_6$.

Posterior vs. Prior vs. Observed mean (disease rate). Examine the posterior distribution of a/b, and compare it to the corresponding prior distribution as well as to the average of y_i/x_i across the six counties.

$$E(\theta_i) = a/b$$

```
a.b <- THETA[burn_in:S, 1]/THETA[burn_in:S,2]
a.b.prior <- rgamma(S-burn_in + 1, 1, 1)/rgamma(S-burn_in + 1,10, 1)
y.x <- mean(y/x)
```

```r
# posterior
plot(density(a.b), xlab = "disease rate",main = "Posterior vs. Prior vs. Observed data (6 counties)", c
abline(v = median(a.b),col = "red", lwd = 2,lty=2)
# prior
abline(v = median(a.b.prior),col="grey", lwd = 2, lty=2)
lines(density(a.b.prior),col = 'grey',lwd = 2)
# observed
abline(v = y.x, col = "blue", lwd = 2, lty=2)
legend('topright', legend=c("posterior.median", "prior.median","y/x.mean"),
       col=c("red", "grey","blue"), lty=2,cex=1,lwd = 3)
```

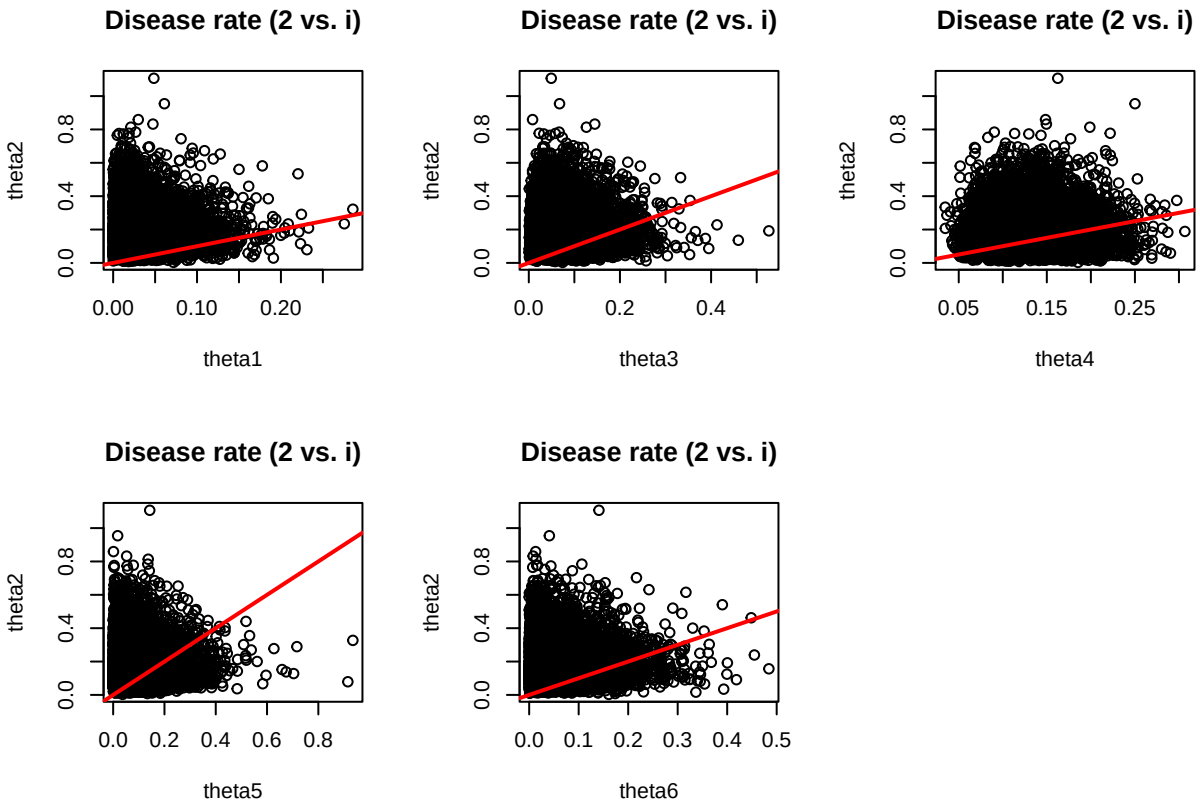


Across 6 counties, posterior median disease rate a/b is smaller than prior median $a.prior/b.prior$ and observed average disease rate y_i/x_i.

Investigate higher disease rates of the second county.

Plot samples of θ_2 versus θ_j for each $j \neq 2$, and draw a 45 degree line on the plot as well.

```r
par(mfrow=c(2,3))
for (j in setdiff(1:6, 2)) {
  plot(theta.post[, j], theta.post[, 2], xlab = paste0("theta",j),
       ylab = "theta2", main="Disease rate (2 vs. i)")
  abline(a = 0, b = 1, col = "red", lwd=2)
}
```



Estimate $Pr(\theta_2 > \theta_j|x, y)$ for each j

$$Pr(\theta_2 > \theta_j|x, y)$$

:

```
sapply(setdiff(1:6, 2), function(j) {
  mean(theta.post[, 2] > theta.post[,j])
})
```

```
## [1] 0.9665 0.8697 0.6656 0.8448 0.9030
```

```
mean(theta.post[, 2] == apply(theta.post, 1, max))
```

Compare yj/xj for each county j vs. Posterior Conclusion ($Pr(\theta_2 = max\{\theta_1, ..., \theta_6\}|x, y)$.)

```
## [1] 0.5783
```

```
y/x
```

```
## [1] 0.03030303 0.21428571 0.07407407 0.13333333 0.08333333 0.05882353
```

- $Pr(\theta_2 = max\{\theta_1, ..., \theta_6\}|x, y) = 0.58$.
- $y_2/x_2 = 0.21$: the highest disease rate among six counties in observed data

Observed data shows the second county has the highest disease rate. Also, by **Bayesian analysis**, posterior estimates show that the second county has high probability to be the county with the largest disease rate.