Chapter 13.4 - 13.5 Comparison of Rates of Two Authors

Jim Albert and Monika Hu

Chapter 13 Case Studies

Introduction

- We collect the counts $\{y_{1i}\}$ of the word "can" in the Federalist Papers authored by Alexander Hamilton and the counts $\{y_{2i}\}$ of "can" in the Federalist Papers authored by James Madison.
- The general problem is to compare the true rates per 1000 words of the two authors.

Negative Binomial Sampling

- The Hamilton counts $y_{11}, ..., y_{1N_1}$ are assumed to be independent Negative Binomial, where y_{1i} is NB(p_{1i}, α_1) with

$$p_{1i} = \frac{\beta_1}{\beta_1 + n_{1i}/1000},$$

and $\{n_{1i}\}$ are the word counts for the Hamilton essays.

- Similarly, the Madison counts $y_{21}, ..., y_{2N_2}$ are assumed to be independent Negative Binomial, where y_{2i} is NB(p_{2i}, α_2) with

$$p_{2i} = \frac{\beta_2}{\beta_2 + n_{2i}/1000},$$

and $\{n_{2i}\}$ are the word counts for the Madison essays.

Comparison

- Focus will be to learn about

$$\mu_M/\mu_H$$

the ratio of the rates (per 1000 words) of use of the word "can" of the two authors, where

$$\mu_M = \alpha_2/\beta_2$$

and

$$\mu_H = \alpha_1/\beta_1$$

Sampling

- Assume that the observed counts of word "can" of the two authors are independent. Moreover, assume that the prior distributions of the parameters (α_1, β_1) and (α_2, β_2) are independent.
- Then the posterior distribution is given, up to an unknown proportionality constant, by

$$\pi(\alpha_1, \beta_1, \alpha_2, \beta_2 \mid \{y_{1i}\}, \{y_{12}\})$$

$$\propto \prod_{k=1}^{2} \left(\prod_{i=1}^{n_{ki}} f(y_{ki} \mid \alpha_k, \beta_k) \pi(\alpha_k, \beta_k) \right).$$

Prior

- Assume that the user has little prior information about the location of the Negative Binomial parameters
- Assume they are independent with each parameter assigned a Gamma prior with parameters 0.001 and 0.001.

R Work

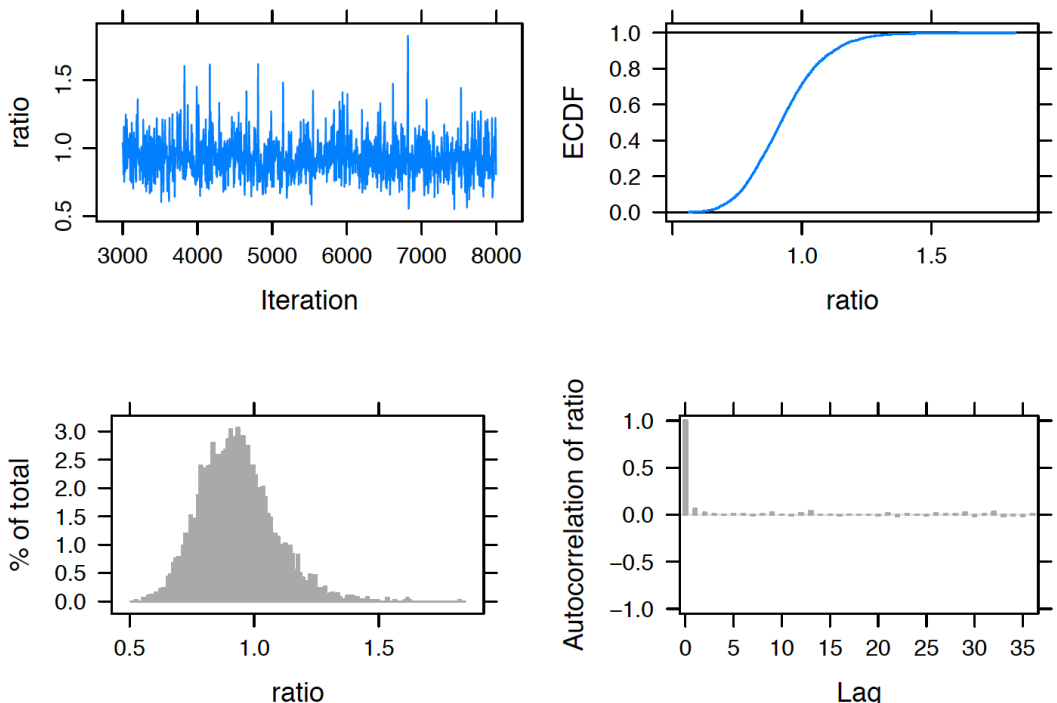
- Posterior sampling is implemented using the JAGS software.
- Model description script is an extension of the previous script for a single Negative Binomial sample.
- The `ratio` parameter is defined to be the ratio of the word rates for the two samples.

JAGS Script

```
modelString = "model{
# Sampling
for(i in 1:N1){
   p1[i] <- beta1 / (beta1 + n1[i] / 1000)
   y1[i] ~ dnegbin(p1[i], alpha1)
}
for(i in 1:N2){
   p2[i] <- beta2 / (beta2 + n2[i] / 1000)
   y2[i] ~ dnegbin(p2[i], alpha2)
}
# Priors
alpha1 ~ dgamma(.001, .001)
beta1 ~ dgamma(.001, .001)
alpha2 ~ dgamma(.001, .001)
beta2 ~ dgamma(.001, .001)
ratio <- (alpha2 / beta2) / (alpha1 / beta1)
}"
```

Learning about ratio of "can" use

- Figure displays MCMC diagnostics for the ratio of "can" rates $R = \mu_M/\mu_H$.
- Posterior median of R is 0.92 and a 95% probability interval for R is found to be (0.71, 1.19).
- There is no significant evidence to conclude that Hamilton and Madison have different rates of the word "can".



Which words distinguish the two authors?

- The previous two-sample analysis was repeated for each of the following words: also, an, any, by, can, from, his, may, of, on, there, this, to, and upon.
- For each word, we focus on inferences about the parameter R, the ratio of mean rates of the particular word by Madison and Hamilton.
- A ratio value of $R > 1$ indicates that Madison was a more frequent user of the word, and a ratio value $R < 1$ indicates that Hamilton used a higher rate of that word.
- Fourteen separate two-sample analyses were conducted and the posterior distributions of R were summarized by posterior medians and 95% probability intervals.

Results

- Figure on the next slide displays the locations of the posterior medians and interval estimates for all of the 14 analyses.
- Intervals that are completely on one side of the value $R = 1$ indicate that one author was more likely to use that particular word.
- One sees that the words upon, to, this, there, any, and an were more likely be used by Hamilton, and the words on, by, and also were more likely be used by Madison.
- The posterior intervals for the remaining words (may, his, from, can, and also) cover the value one, and so one cannot say from these data that one author was more likely to use those particular words.

Figure of ratio of rates for 14 words