AGEING POPULATION GROWTH ANALYSIS USING BAYESIAN INFERENCE

Li Sun

1. Objective

The world is getting older by the minute. It is set to become one of the most consequential social alterations in the age of humankind. In this project, a data set (from data.worldbank.org) containing population ages 65 and above as a ratio of the total population for 269 countries was analyzed. After omitting the empty observations, there were a total of 237 countries left in the final data model. The comparison of ageing population ratio changes globally between the first 2 years in the data set, which were the years of 1960 and 1961, and between the latest 2 years 2016 and 2017 was simulated. Since the model parameters were random variables, Bayesian inference as a feasible approach was implemented. Furthermore, in Bayesian statistics, since Markov chain Monte Carlo (MCMC) methods has been a key step in making it possible to compute large hierarchical models, MCMC method was utilized in the work. At the same time, Metropolis-Hastings algorithm, a specific implementation of MCMC, was performed to compute the distribution over the parameters.

The questions were formulated as:

• Is the mean of ageing population ratio in 1961 greater than the mean of ageing population ratio in 1960? If so, how much?

• Is the mean of ageing population ratio in year 2017 greater than the mean of ageing population ratio in year 2016? If so, how much?

• Is the mean of ageing population ratio in year 2017 greater than the mean of ageing population ratio in year 1960? If so, how much?

2. Computational Method Introduction

2.1 Bayesian inference

Bayesian inference is closely related to probability. It manipulates priors, evidence, and likelihood to computer posterior. It consists of the following elements:

- prior: $P(\theta)$

  represents how likely the parameter (θ) is

- likelihood: $P(D \mid \theta)$

  represents how likely is data (D) for given parameter (θ)

- evidence: P(D)

  represents how likely data (D) is

- posterior: P(θ | D)

  represents the most likely distribution of parameter (θ)of the model explaining the data (D)

  P(θ | D) = $\frac{P(\theta)*P(D \mid \theta)}{P(D)}$

  posterior P(θ | D) is proportional to prior P(θ) * likelihood P(D | θ)

2.2 Markov Chain Monte Carlo (MCMC)

MCMC allows us to draw samples from a distribution even if we cannot compute it. It can be used to compute the distribution over the parameters given a set of observations and a prior belief.

2.3 Metropolis-Hastings Algorithm

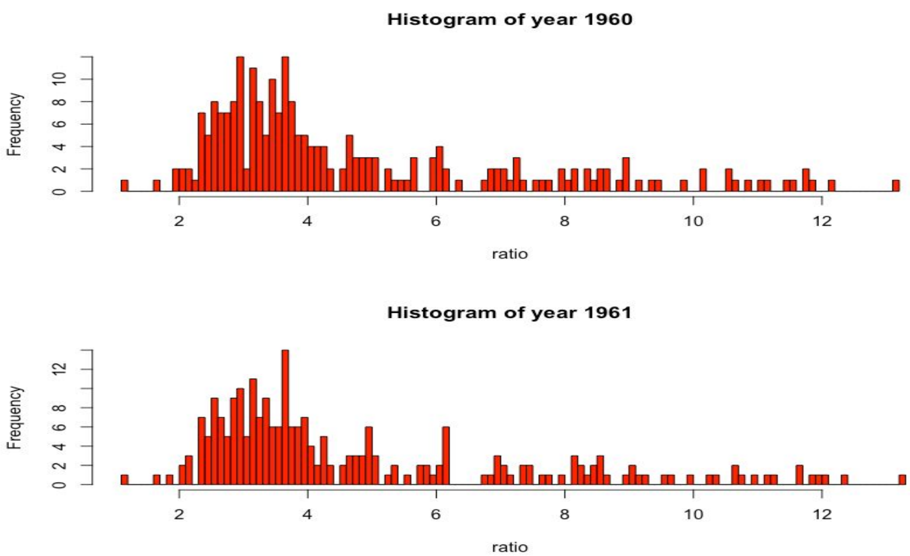
Metropolis-Hastings is a specific implementation of MCMC. It works well in high dimensional spaces in sampling and rejection sampling.The Metropolis-Hastings algorithm does the following:

- given
  1. f, the PDF of the distribution to sample from
  2. Q, the transition model
  3. θ_0, a first guess for θ
  4. $\theta = \theta_0$
- for n iterations
  1. *p = f(D|$\Theta = \theta$)P(θ)*
  2. *$\theta' = Q(\theta_i)$*
  3. p' = f($D|\Theta = \theta')P(\theta')$
  4. ratio = p' / p
  5. generate a uniform random number r in [0,1]
  6. if r < ratio, set $\theta_i = \theta'$

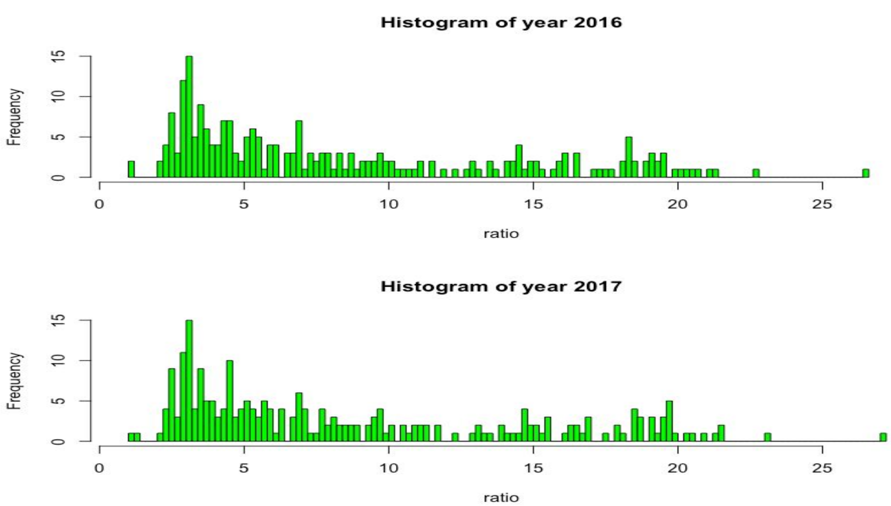
3. Data Analysis

3.1 Computational method implementation

The histograms of ratios in 1960 and 1961 are given below:



The histograms of ratios in 2016 and 2017 are given below:



The conclusion is not so obvious from histograms. But it is not hard to see that all of them are right-skewed distribution. I assumed that the ratio in each year followed gamma distribution Γ *(a,b)*, here *a* is shape parameter, and *b* is a scale parameter. The ratio in pre-year (1960 or 2016) comes from $\Gamma(a_1,\ b_1)$ and the tatio in pos-year (1961 or 2017) comes from $\Gamma(a_2,\ b_2)$. So the parameter vector is

$\theta = (a_1,\ b_1 a_2, b_2)$. One important point that has to be emphasized is that the mean of a gamma distribution is *a * b*.

I constructed a likelihood function and a prior function, and then multiplied them together to get posterior *P(θ|Data)*. In my case, multiplying hundreds of small values caused an underflow in the system memory. In order to keep numerical stability, I used ***log*** on both the prior and the likelihood functions.

3.1.1 Likelihood function

Let x denote the ratio in pre-year and let y denote the ratio in pos-year. Since x_i and y_j are independent for *i = 1,2, …, 237* and *j = 1,2,..., 237*, the likelihood function is written by the formula

- likelihood: P(D | θ)

$$P(D|\theta) = P(x|\theta) * P(y|\theta)$$

$$= \prod_{i=1}^{237} \Gamma(x_i \mid a_1, b_1) \prod_{j=1}^{237} \Gamma(y_j \mid a_2, b_2) = \prod_{i=1}^{237} \left(\frac{1}{{b_1}^{a_1} \cdot \Gamma(a_1)} \cdot x^{a_1 - 1} \cdot e^{-\frac{x}{b_1}} \right) \prod_{j=1}^{237} \left(\frac{1}{{b_2}^{a_2} \cdot \Gamma(a_2)} \cdot x^{a_2 - 1} \cdot e^{-\frac{x}{b_2}} \right)$$

- log(*Likelihood):*

log(P(D|θ)) = log (P(x|θ) P(y|θ))*

*= sum((a1-1)*log(x_i) - (1/b1)*pre - a1*log(b1) - log(gamma(a1))) +*

*sum((a2-1)*log(y_j) - (1/b2)*pre - a2*log(b2) - log(gamma(a2)))*

3.1.2 Prior function

As for the decision on a prior density for $\theta = (a_1, b_1, a_2, b_2)$, the basic thinking is simply to choose priors for the *mu1* = $a_1 * b_1$ and *mu2* = $a_2 * b_2$ that does not seem to prejudge which is big, and to choose priors that are sort of uniformish over all even remotely plausible values, with the aim of letting the data produce any interesting features in the posterior distribution. I chose the distribution of all 4 parameters as normal distribution, that is $a_1 \sim N(a_{01}, a_{01})$, $b_1 \sim N(b_{01}, b_{01})$, $a_2 \sim N(a_{02}, a_{02})$ and $b_2 \sim N(b_{02}, b_{02})$. I chose the distributions like this because I wanted the distribution as flat as possible, otherwise the prior would control the posterior.

Then the prior is given by

- prior: P(θ)

$$P(\theta) = P(a_1, b_1, a_2, b_2) = P(a_1) * P(b_1) * P(a_2) * P(b_2)$$

$$= \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi} \cdot a_{0i}} e^{-\frac{1}{2 \cdot a_{0i}^2}(a_i - a_{0i})^2} \cdot \prod_{j=1}^{2} \frac{1}{\sqrt{2\pi} \cdot b_{0j}} e^{-\frac{1}{2 b_{0j}^2}(b_j - b_{0j})}$$

- log(prior: P(θ))

  log(P(θ)) = log (P(a_1)*P(b_1)*P(a_2)* P(b_2))

  = -log(a01*sqrt(2*pi)) - ((a1-a01)^2/(2*a01^2))
  -log(b01*sqrt(2*pi)) - ((b1-b01)^2/(2*b01^2))
  -log(a02*sqrt(2*pi)) - ((a2-a02)^2/(2*a02^2))
  -log(b02*sqrt(2*pi)) - ((b2-b02)^2/(2*b02^2))

3.1.3 Posterior function

Next, posterior is obtained by multiplying prior and likelihood. And log posterior is the sum of the log prior and log likelihood.

- posterior P(θ | D)
  P(θ | D) = prior P(θ) * likelihood P(D | θ)

- log(posterior P(θ | D))
  Log (P(θ | D)) = log (P(θ)) + log(P(D | θ))

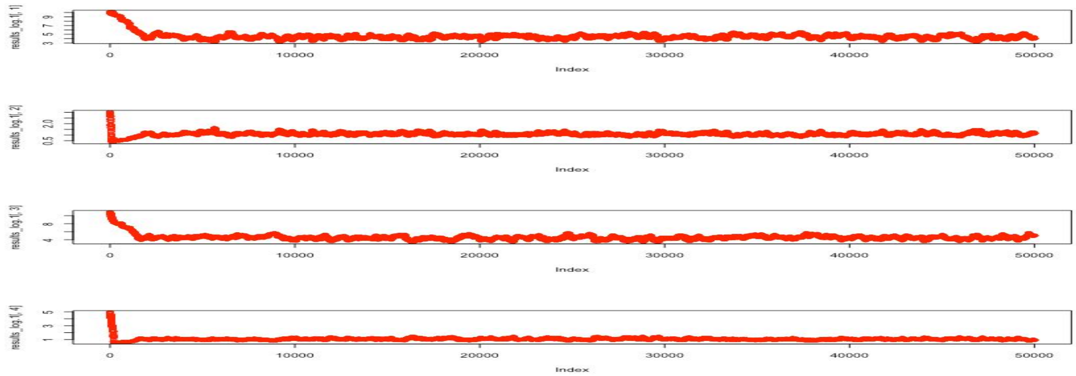
3.1.4 Metropolis-Hastings Algorithm manipulation

In the project, and in both cases, the initial parameters were randomly assigned as $a_{01}=10$, $b_{01}=3$, $a_{02}=11$, *and* $b_{02}=5$. A Markov chain using Metropolis for 50,000 iterations was performed to simulate a sample. Given the current state, a way to propose a ‘candidata’ move has to be decided. Then the log posterior of candidate move and the log posterior of current state are compared. Due to the log algorithm, exponentiation is used to compare to the random number in acceptance and rejection samplings. The results are recorded in a big matrix with 4 columns and 50000 rows. The first row is the starting value and each successive row records the next θ as the chain runs.

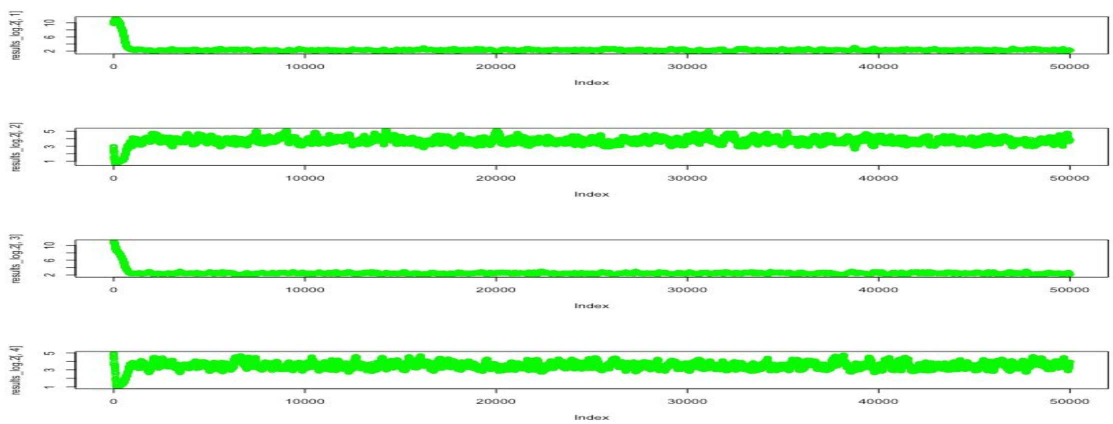
3.2 Result interpretation

3.2.1 data Analysis and interpretation

Now we have gotten a bunch of parameter vectors. Let’s first take a look at how the chain ran. The traces for parameters in 1960 and 1961 are given below.

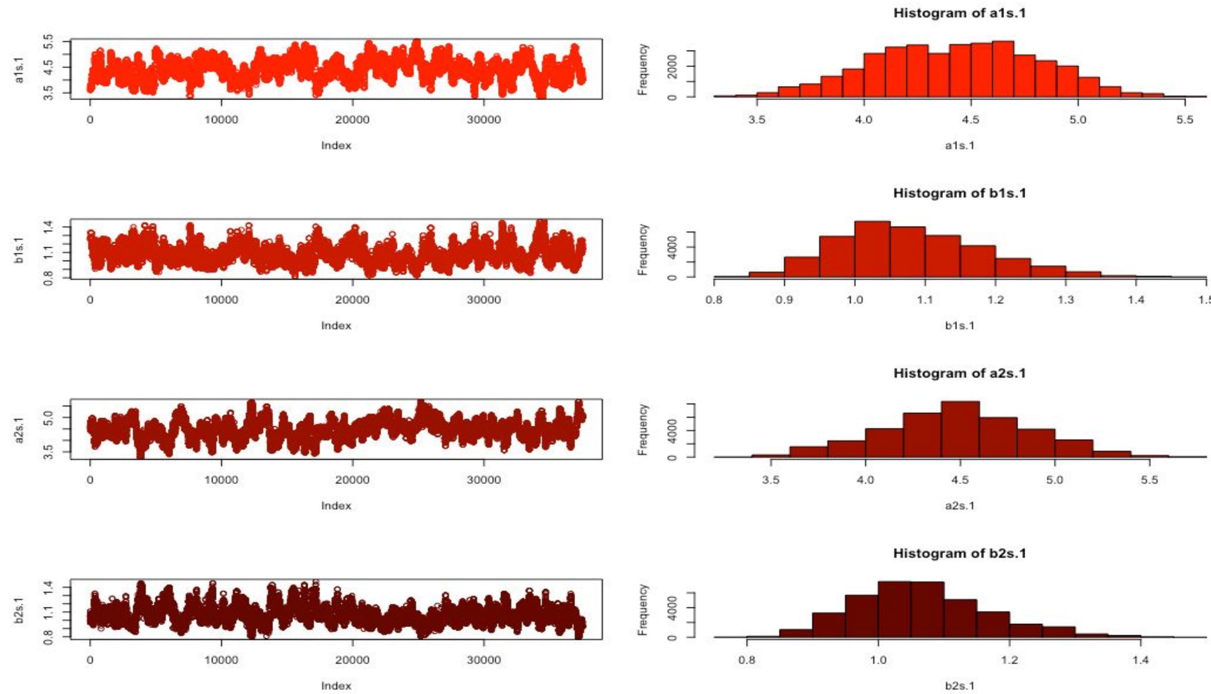


The traces for parameters in 2016 and 2017 are given below:

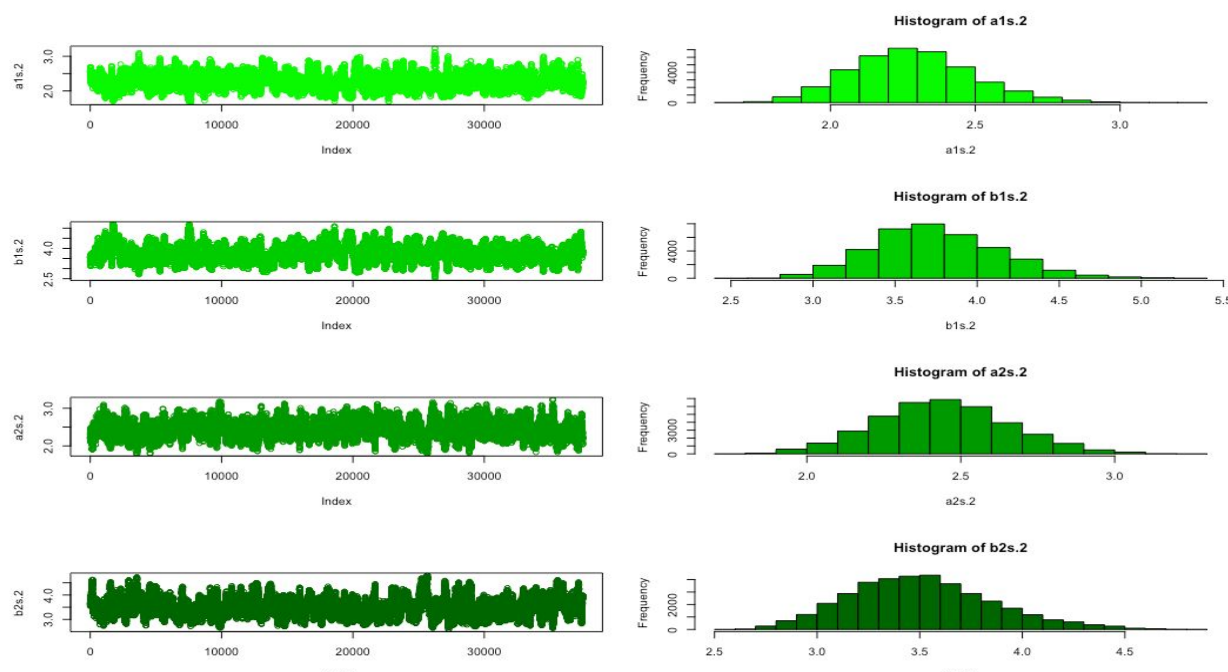


It looks like the beginning iterations may be noticeably influenced by the starting values. Then I threw away the first 25% of iterations as ‘burn-in’ and let the remaining be the new results. In this way, the inferences were based on the new results instead of the whole results matrix. The new MCMC sampling without ‘burn in’ is shown below. The parameter histograms indicate that the normality assumption of initial parameters for prior function is reasonable and acceptable.

The traces of parameters without ‘burn-in’ and their corresponding histograms in 1960 and 1961 are given by

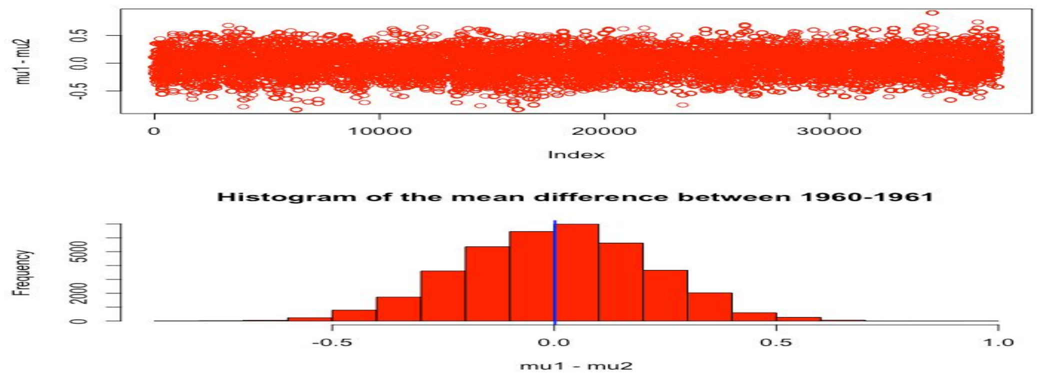


The traces of parameters without 'burn-in' data and their corresponding histograms in 2016 and 2017 are given by

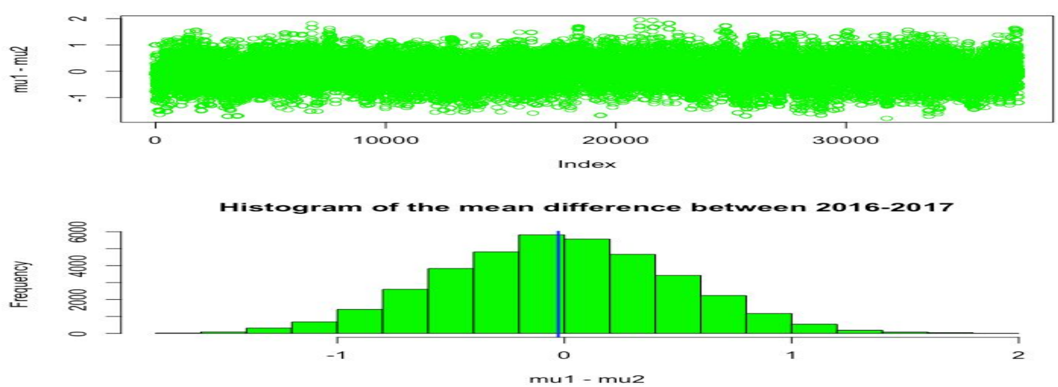


Based on the estimated posterior, the mean difference between the pre-year and pos-year were computed.

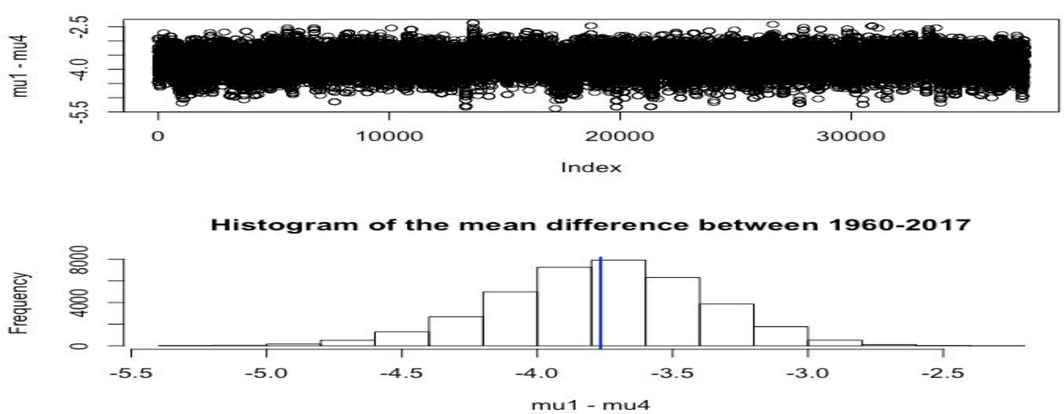
The plot and histogram of the mean difference between 1960 and 1961 are posted as



The plot and histogram of the mean difference between 2016 and 2017 are posted as



The plot and histogram of the mean difference between 1960 and 2017 are posted as



The estimated posterior probabilities that *mu1 - mu2 < 0* are 0.487, 0.522 and 1 for years 1960 and 1961, years 2016 and 2017, and years 1960 and 2017 respectively. Thus the probability that the mean of ageing population ratio in 1961 higher than 1960 is slightly lower than 0.5. In contrast, the probability that the mean ageing population ratio in 2017 higher than 2016 is slightly greater than 0.5. There is no surprise that the mean of ageing population ratio would increase dramatically during the past 57 years from 1960 to 2017, however it is still astonishing that the probability is 100% for sure that the mean of ageing population ratio in 2017 is higher than 1960.

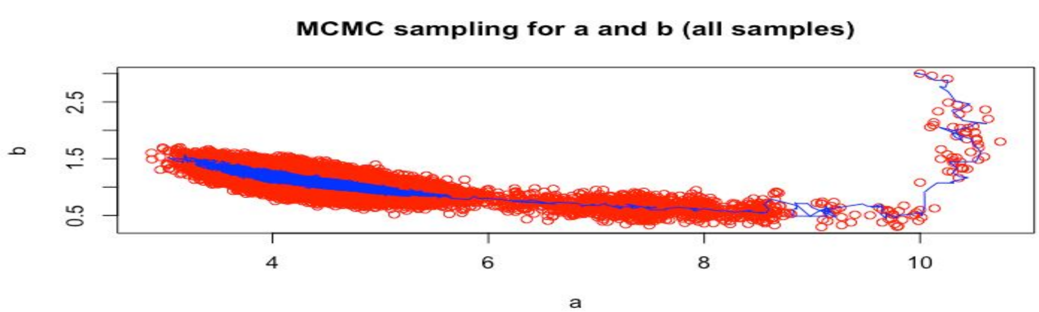
.

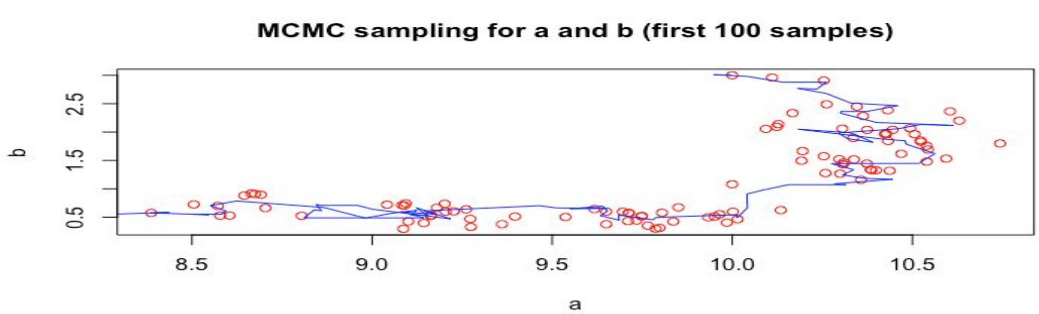
3.2.2 Example analysis for year 1960

In this project, I explored more detail on the simulated parameter data in 1960. The starting value for shape *a* is *10*, and for scale *b* is *3*. The MCMC Metropolis-Hastings algorithm accepted 7598 pairs of samples and rejected 42403 pairs of samples. The average of value without burn-in for shape *a* is *4.39136*, and for scale *b* is *1.08786,* which are pretty far off the initial values.

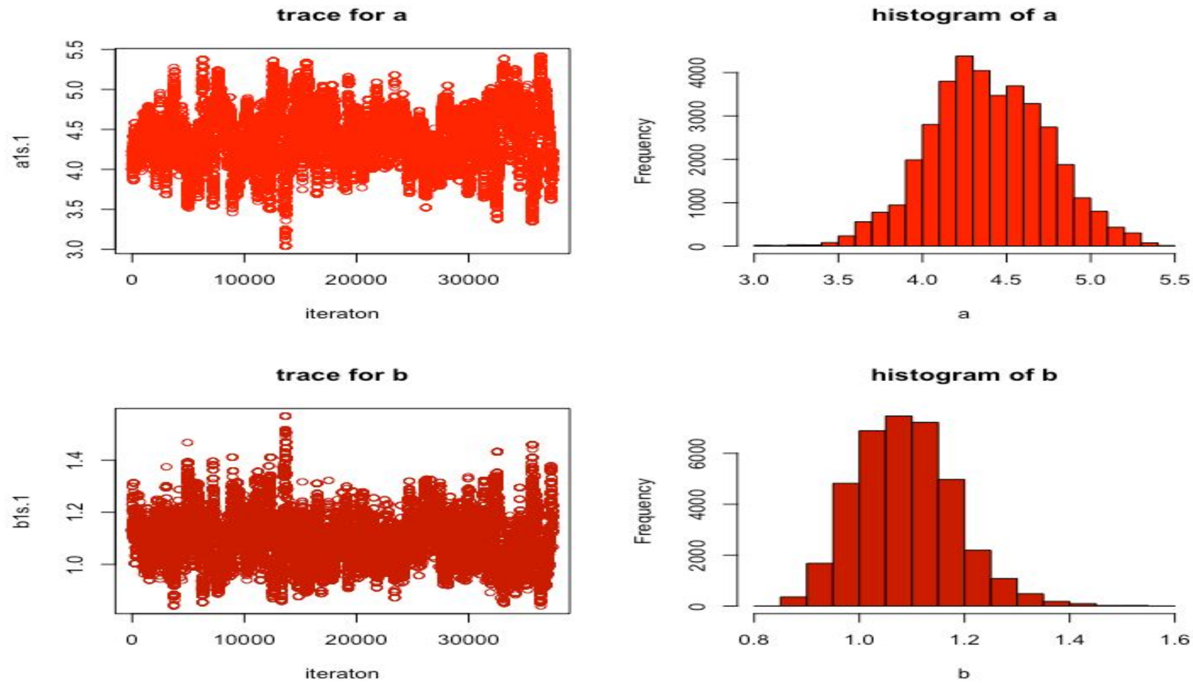
The plots of MCMC sampling for a and b with Metropolis-Hastings illustrate how the algorithm worked its way to the parameters. In these plots, the accepted samples are presented in blue lines, and rejected samples are marked as red dots.

As we can see from the figures, the algorithm converges to the zone around [*a = 4.4*, *b = 1.1*].

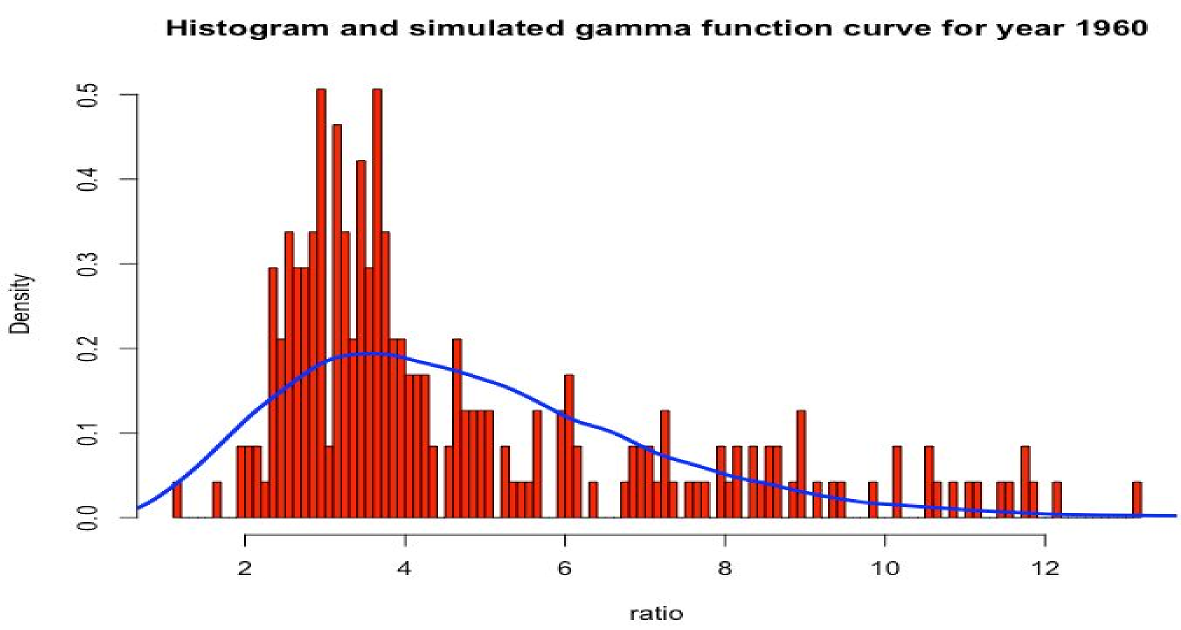




The traces of a and b, and the histogram of the traces are shown below.



The prediction of the distribution of ageing population ratio in year 1960 with a gamma distribution $\Gamma(a = 4.39136,\ b = 1.08786)$ are graphed below. It indicates that proposal gamma distribution works fine in this case.



4. Conclusion

According to our discussion above, we get the conclusion that

- The mean of ageing population ratio in 1961 is slightly less than the mean of ageing population ratio in 1960. The probability of *mean(1961) > mean(1960)* is 0.49.

- The mean of ageing population ratio in 2017 is slightly greater than the mean of ageing population ratio in 2016. The probability of *mean(2017) > mean(2016)* is 0.52.

- The mean of ageing population ratio in 2017 is definitely greater than the mean of ageing population ratio in 1960. The probability of *mean(2017) > mean(1960)* is 1.

Thus overall, the change in the ratio of the population ages 65 and above over the total population between 1961 and 1960 decreases slightly, and the change in ratio increases slightly from 2017 to 2016. When taking a long term view, the change in the ratio between 2017 and 1960 increases certainly.