Bayesian Analysis on Mortality Rate of Covid-19 in California

Yuxin Zhang

Abstract

This paper presents the second analysis of the Covid-19 data in the state of California, with 58 counties. The data includes population, total cases confirmed and numbers of deaths in each county. The focuses of interest are on the rate of mortality, and infection rate corresponding to each county. Using Bayesian analysis approaches, we propose two models in this paper, compare their prediction and assess the validity of each model

Two proposed models in this report display similar results for posterior prediction of number of deaths for each county, which also captures the trend of our real data. Los Angeles, San Diego and Riverside are among the top 3 highest death numbers among all other counties.

Key Words: Bayesian, hierarchical, MCMC, Poisson

1. Introduction

With the dramatic spread in the novel Covid-19, the World Health Organization (WHO) and many countries have been tracking the latest data on number of confirmed cases over the past few months. As more number of data sets become accessible, data scientists around the world are conducting analyses, building up algorithms and making prediction. It is inevitably that Bayesian inference will play a big part of it. Bayesian modeling can be applied to efficiently analyze the growth of Covid-19.

The data set contains data of Covid-19 in 58 counties in California as of April 13,2020. There are four columns: County name; Number of cases confirmed in each city, denoted as n_i for the $i-th$ county, number of deaths, denoted as y_i and the population of the county. The project aims to analyze different Bayesian models and to understand under each model if there are any differences between counties regarding Covid-19 influence.

Data	sd	Range
Population	1285.656	(0,9420)
Cases	1466643	(2930,10039108)

Table 1: Variability Measurement for Total Population and Total Cases

1.1 Understanding the variability of dataset

Understanding and characterizing variability in samples is an important part of data analysis. Variability can be measured with several different statistics, such as range, standard deviation, quantile, and etc. Standard deviation measures the spread of a data distribution. The more spread out a data distribution is, the greater its standard deviation Both standard deviation for total population and total cases are extremely high, given the extreme values existing in the dataset. Since range tends to increase with sample size, and because it is highly sensitive to outliers, it is not a desirable measure of variation in the Covid-19 dataset. Table[1] displays the variability of total population and cases in the dataset.

In regarding the measurement of variation, the box plot is a standardized way of displaying the distribution of data. Figure[1] shows two boxplots: the left is the boxplot of log of total cases and the right plot shows the distribution of log of total population. The reason that the data are log transformed is because of the extreme values that make the boxplots highly skewed. This can be valuable for making patterns in the data

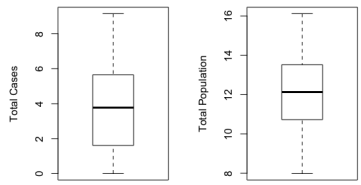

Figure 1: Boxplots of total cases (after logarithm, shown in left panel) and of total population rate (after logarithm, shown in right panel)

more interpretable. Even after the transformation, total cases are still relatively more left skewed comparing with the total population. In addition, Figure[2] shows that as the total population increases, the total cases also increases, a potential linear relationship.

Figure 2: Scatterplot of Total Cases vs Total Population

2. Inference with Poisson and Gamma Model

In the first model, we propose a Poisson model for the number of cases denoted as n_i for the i-th county. We assume that for each county, the mean number of cases confirmed is equal to 20% of the population. The proposed prior distribution is Gamma with parameter (α, β).

2.1 Model Specification

The model 1 is displayed below:

$$n_i \sim Pois(\lambda_i c_i/10^3), \lambda_i \sim Ga(\alpha, \beta),$$

$$p(\alpha, \beta) = Ga(a_\alpha, b_\alpha)Ga(a_\beta, b_\beta)$$

2.1.1 Hyperprior

From our assumption about 20% case rate, the $\lambda_i = 200$ with the following logic:

$$E(n_i) = \frac{c_i \lambda_i}{1000} = 20\% c_i$$

which returns $\lambda_i = 200$.
Since λ follows a Gamma distribution

$$E(\lambda_i) = \frac{\alpha}{\beta} = 200$$

For the simplicity, we assume $\alpha = 200, \beta = 1$, then assume $a_\alpha = 200, b_\alpha = 1$ and $a_\beta = 0.5, b_\beta = 0.2$

s

2.1.2 Full Joint Distribution

With the hyperparameters set up, we can obtain posterior distribution and to obtain samples of λ, α, β. The joint posterior distribution is multiplication of gamma distribution and poisson distribution shown in equation

$$p(\lambda, \alpha, \beta) = p(\lambda|\alpha, \beta, n)p(\alpha, \beta|n)$$

$$\propto p(\alpha, \beta)p(\lambda|\alpha, \beta)p(n|\lambda, \alpha, \beta)$$

$$\propto Ga(a_\alpha, b_\alpha)Ga(a_\beta, b_\beta)$$

$$* \prod_{i=1}^{58} e^{-\lambda_i c_i/1000} (\frac{\lambda_i c_i}{10^3})^{n_i}$$

Using factorization, we obtain the sample from the full conditional distribution of λ

$$p(\lambda|\alpha, \beta, n) = \prod_{i=1}^{58} \frac{(\beta + c_i/10^3)^{\alpha+n_i}}{\Gamma(\alpha + n_i)}$$

$$* \gamma_i^{\alpha+n_i-1} e^{-\lambda_i(\beta+c_i 0^3)}$$

Then, by conditional formula,

$$p(\alpha, \beta|n) = \frac{p(\lambda, \alpha, \beta)}{p(\lambda|\alpha, \beta, n)}$$

$$\propto \alpha^{a_\alpha - 1} e^{-b_\alpha \alpha} \beta^{a_\beta - 1} e^{-b_\beta \beta} \prod_{i=1}^{58} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{(c_i/10^3)^{n_i}}{(n_i)!}$$

$$* \prod_{i=1}^{58} \frac{(\beta + c_i/10^3)^{\alpha + n_i}}{\Gamma(\alpha + n_i)}$$

2.2 Sample Based Approach

After deriving the necessary distribution, we used factorization to fit our model. We first sample from hyperpriors $p(\alpha, \beta|n)$ to draw 10000 samples of α, β. Instead of using its direct form, we perform log transformation for both α, β.We use Sampling Importance Resampling algorithm for sampling the parameters.

2.2.1 Sampling Importance Resampling

We simulate 10000 draws from the proposed prior model $p(\alpha, \beta|n)$ with $a_\alpha = 200, b_\alpha = 1$ and $a_\beta = 0.5, b_\beta = 0.2$. We denote

$$\theta_1 = log(\alpha), \theta_2 = log(\beta)$$

. The way that Sampling Importance Resampling (SIR) works is that for m number of draws $\theta_i^1, ...\theta_i^m$, we compute the weights

$$w(\theta_i^j) = g(\theta_i^j|y)/p(\theta_i^j), i = 1, 2$$

The convert the weights to probabilities

$$p_i^j = \frac{w(\theta_i^j)}{\sum_{j=1}^m w(\theta_i^j)}$$

The algorithm can be implemented using sir built-in function in R. Given the sample of θ (after converting back to α, β by exponentiating) which would be approximately distributed according the the distribution g, we complete the Sampling Importance Resampling steps. To explore the properties of the joint distribution, we plot the histogram for the samples of α, β shown in Figure[3]

2.2.2 Full Conditional Distribution

From the sampled α, β, we obtain the samples of λ, and $p(\lambda|\alpha, \beta) \sim Ga(\sum_{i=1}^{58} n_i + \alpha, \sum_{i=1}^{58} c_i/10^3 + \beta)$

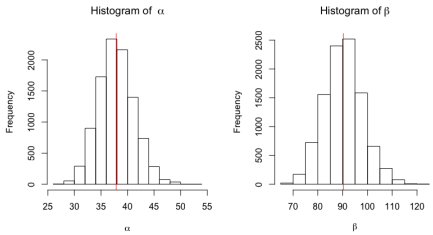

Figure 3: The histogram of sampled α(left panel), and sampled β (right panel). The red line show the mean of the distribution

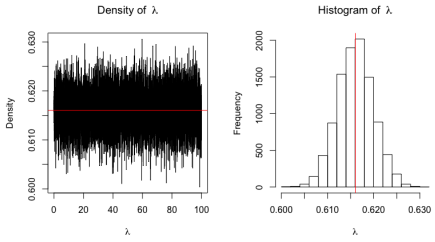

Figure 4: The density plot of sampled λ(left panel), and histogram (right panel). The red lines in both plots show the mean of the distribution

2.3 Analysis of Results

As we obtain the samples of our parameters of interest, we can start evaluate the model fitness, obtaining samples from the posterior predictive distribution n_i. Figure[5] displays headshot of the predicted total number of cases, the average number of cases in the given time interval for each county.

County	Expected rate	Model 1 Predicted cases	Total cases
LA	0.934	10040	9420
SD	0.550	3338	1847
RIV	0.712	2424	1751
SCL	0.840	1939	1666
OC	0.402	3191	1283
SBD	0.451	2157	977
SF	1.020	885	957
ALA	0.527	1664	886
SAC	0.479	1531	739
SM	0.857	772	701
CC	0.477	1148	552
KER	0.492	893	446

Figure 5: The headshot table of predicted death for few counties

3. Inference with Hierarchical Model and Model 1

3.1 Model Specification

Consider a hierarchical model

$$y_i \sim Bin(n_i, \theta_i), \theta_i \sim Be(\mu\tau, (1-\mu)\tau)$$

$$p(\mu, \tau) = (\mu(1-\mu)(1+\tau)^2)^{-1}$$

We can factorize the joint distribution

$$p(\theta, \mu, \tau|y) = p(\theta|\mu, \tau, y)p(\mu, \tau|y)$$

By using the factorization, we can draw samples of μ, τ from $p(\mu, \tau|y)$ and then plug the samples to obtain samples from $p(\theta|\mu, \tau)$ to get different mortality for each county.

3.1.1 Hypterprior

First, we find the joint distribution for $(\mu, \tau|y)$. We observe that

$$p(\mu, \tau|y) \propto p(y|\mu, \tau)p(\mu, \tau)$$

$$p(\mu, \tau|y) = \prod_{i=1}^{58} \binom{n_i}{y_i}$$

$$* \frac{B(\mu\tau + y_i, \tau(1-\mu) + n_i - y_i)}{B(\mu\tau, \tau(1-\mu)) * \mu(1-\mu)(1+\tau)^2}$$

3.1.2 Full Conditional

Next, to solve the conditional posterior for θ, we have:

$$p(\theta|\mu, \tau, y) = \prod_{i=1}^{58} p(\theta_i|\mu, \tau, y_i)$$

$$\propto \prod_{i=1}^{58} Be(\theta_i|\mu\tau + y_i, \tau(1-\mu) + n_i - y_i)$$

3.1.3 Sample Based Approach

With the model specified, we can draw samples of μ, τ using the same Sampling Importance Resampling (SIR) described in the previous model. We denote $\theta_1 = log(\mu), \theta_2 = log(\tau)$. Using SIR, Laplace mode of μ, τ = (-3.520815,5.455601) and the Laplace variance = $\begin{bmatrix} 0.009 & -0.0142 \\ -0.0142 & 0.2635 \end{bmatrix}$.We draw 10000 samples and illustrate their distribution in Figure[6]

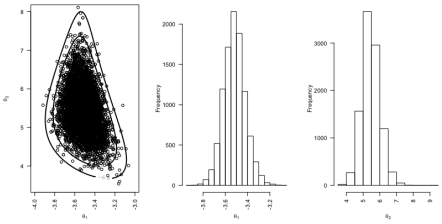

Figure 6: The contour plot of θ_1, θ_2 (left panel), the histogram of sampled θ_1 (middle panel) and the histogram of sampled θ_2(right panel)

Then, we transform back μ, τ by exponentiating them. Plugging the samples into the full conditional distribution of θ, we obtain 10000 samples of θ. Figure[7] shows the histogram of posterior samples. The mortality rate has a mean of 2.98%, median of 2.98% and standard deviation of 0.0011.

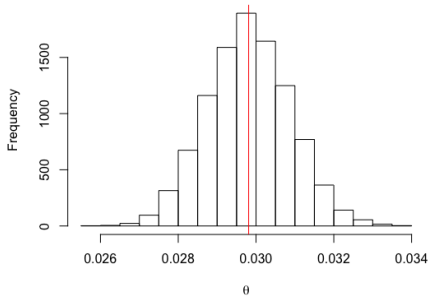

Figure 7: The histogram of sampled θ, the red line shows the mean of posterior samples

3.2 Analysis of Results

Now we have both samples of number of cases: n from Model 1 and mortality rate: θ obtained through the hierarchical model. To understand if our model fits well, we look at the posterior predictive distribution of the number of deaths y_i for each county. From our hierarchical model, the number of deaths for each county $(i = 1, ...58)$ has the distribution of binomial with parameters of n_i, θ_i. Therefore, we simulate 10000 samples of

y_i from the posterior samples. Figure[8] shows the mean of posterior samples for each county. The prediction tend to capture the true data fairly well given that Los Angeles has the highest death numbers among all other counties. However, if we take the population of each county into consideration, Los Angeles is no longer the county with the highest posterior mean, but San Jose is. Figure[9] shows the mean of posterior deaths per 1000 habitants for each county.

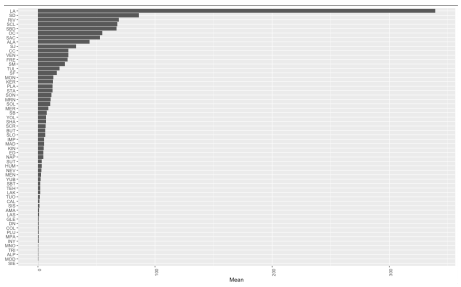

Figure 8: The barplot of posterior mean of deaths of each county, sorted by mean

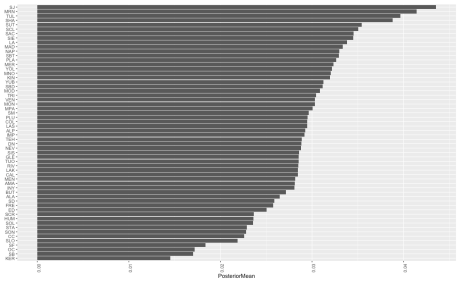

Figure 9: The barplot of posterior mean of deaths per 1000 habitants of each county, sorted by mean

4. Inference with Binomial, Poisson, Beta, Gamma Models

4.1 Model Specification

Now consider a modification of the models where the rate of infection depends on the mortality:

$$f(y_i, n_i|\theta_i, \lambda) = f(y_i|n_i, \theta_i, \lambda) f(n_i|\theta_i, \lambda)$$

$$= Bin(y_i|n_i, \theta_i) Pois(n_i|\theta_i \lambda c_i/10^3)$$

for $\theta_i \epsilon (0,1)$, assume $\theta_i \sim Be(\theta_i|\mu, \tau, (1-\mu)\tau)$ with prior defined as

$$p(\mu, \tau) = (\mu(1-\mu)(1+\tau)^2)^{-1}$$

$\mu \epsilon (0,1), \tau > 0$ and $\lambda \sim Ga(\lambda|a,b)$

4.1.1 Full Joint Distribution

With a complex hierarchical model like this, to obtain samples of our parameters of interest, we first write out the full joint distribution between parameters:

$$p(\lambda, \mu, \tau, \theta|n, y)$$

$$\propto \prod_{i=1}^{58} [\theta_i^{y_i}(1-\theta_i)^{n_i - y_i}$$

$$*(\frac{\theta_i \lambda c_i}{10^3})^{n_i} * \binom{n_i}{y_i} * \frac{1}{(n_i)!}$$

$$*e^{\frac{-\theta_i \lambda c_i}{10^3}} * \frac{\Gamma(\mu)}{\Gamma(\mu\tau)\Gamma(\tau(1-\mu))}]$$

$$*\frac{1}{\mu(1-\mu)(1+\tau)^2} * \lambda^{a-1} e^{-b\lambda}$$

4.1.2 Full Conditionals of Parameters

Now that we specify the hierarchical model and the full joint distribution, the full conditionals for the random parameters as follows. The full conditional for θ_i

$$p(\theta_i|\lambda, \mu, \tau) \propto \theta_i^{y_i + n_i + \mu\tau - 1}(1-\theta_i)^{n_i - y_i + \tau(1-\mu) - 1}$$

$$*e^{-\frac{\theta_i \lambda c_i}{10^3}}$$

The full conditional for λ

$$p(\lambda|\theta, \mu, \tau) \propto \lambda^{\sum_{i=1}^{58} + a - 1} e^{-(b + \frac{\sum_{i=1}^{58} \theta_i c_i}{10^3})\lambda}$$

So $\lambda|\theta, \mu, \tau \sim Ga(\sum_{i=1}^{58} + a, b + \frac{\sum_{i=1}^{58} \theta_i c_i}{10^3})$

The joint posterior distribution for μ, τ

$$p(\mu, \tau|y) \propto p(y|\mu, \tau) p(\mu, \tau)$$

$$p(\mu, \tau|y) = \prod_{i=1}^{58} \frac{\Gamma(\tau)}{\Gamma(\mu\tau)\Gamma(\tau(1-\mu))}$$

$$*\theta_i^{\mu\tau}(1-\theta_i)^{\tau(1-\mu)} * \frac{1}{\mu(1-\mu)(1+\tau)^2}$$

For the three distributions, only the full conditional for λ is a closed form, others do not have known distribution. In the following section, we discuss the sampling method - MCMC we use to sample from their posterior distribution.

4.2 Sample Based Approach

4.2.1 Methodology

After deriving the full conditionals, we notice that each full conditional also contains other parameter of interests. Therefore, to ensure the accuracy of the sampling, we need to alternate between parameters. We sample every posterior distribution twice with the updated parameters

For the setup, we assume $a = b = 0.001$. Later, we will discuss how the choice of a, b can influence the results. Since λ has a closed form, we first draw 10000 samples of λ. We notice that the full conditional contains one parameter of interest: the mortality rate, denoted as θ_i.The support set of $\theta_i \epsilon (0, 1)$. We initiate a $\theta = 0.2$. To obtain 10000 samples of λ_i, we use Metropolis-Hastings with a random walk proposal. Then we plug the sampled λ_i in the full conditionals of θ_i. We assume the $\mu = \tau = 1/2$ as their initial values. Again, we obtain 10000 samples of θ_i using Metropolis-Hastings with Gibbs sampler. Next step is to sample μ, τ with the sampled θ_i, λ_i. μ, τ is sampled using SIR as the previous model does. Last but not the least, repeat the whole process by updating samples at each chain.

4.3 Analysis of Results

Again, we want to check the fitness of our model by checking the posterior predictive distribution for number of deaths denoted as y_i for each county. The highest death numbers again is Los Angeles, same with the previous model. The top three are Los Angeles, San Diego and Riverside County. The two models display similar trend for the numbers of death for each county. However, if we look at the predictive distribution of the number of death per 1000 habitants for each county, the highest one is no longer San Jose. instead it's the Mono County. In addition, the plot doesn't display a similar trend as the previous model and shows a relatively smaller ratios for each county. We have $y_i \sim Bin(y_i|n_i, \theta_i)$

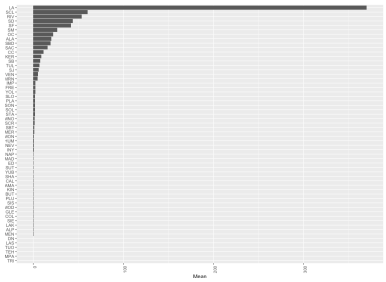

Figure 10: The barplot of posterior mean of deaths for each county, sorted by mean, using Model 2

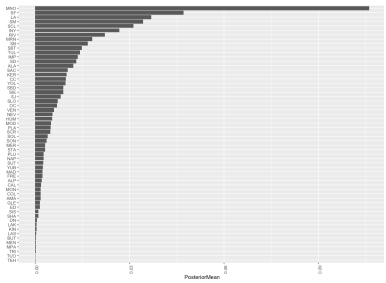

Figure 11: The barplot of posterior mean of deaths per 1000 habitants for each county, sorted by mean, using Model 2

Conclusion In this report, our goal is to understand the mortality rate of Covid-19 for each county. We build two models and analyze if the model could accurately predict the data. The first model consider the number of infection is independent of the mortality rate, but the second model does consider the rate of infection depends on mortality rate. Both models have similar results for predicting number of deaths for each county, but show different results for predicting the number of deaths per 1000 habitants. This result is unexpected given that the population is not a random number. Further investigation on why models behave different should be done. Future direction can be proposed to collect more rel-

evant data such as the healthcare enrollment rate, gender, race and more for each county. In addition, as Covid-19 quickly spreads everyday and the data has been updated nearly every second, this dataset might be outdated as we finish performing the analysis. So how to quickly replicate the analysis as soon as the data is updated is also worth to discuss.