STAT207 Take Home Report

Meltem Ozcan *

Abstract

In this paper, we analyze a data set containing information on the incidence of and mortality due to COVID-19 at the county level in California in order to make predictions regarding the number of deaths due to COVID-19 in the future. We perform an exploratory analysis of the data and then consider three models: a binomial model, a beta-binomial model, and a hierarchical model. Using these models, we make predictions regarding the number of deaths in CA due to COVID-19 in the future, and we compare and discuss the results.

Key Words: COVID19, Bayesian, Hierarchical Models, Beta-Binomial, Rejection Sampling

1. Introduction

COVID-19, a pneumonia-like infectious disease that transmits through droplets in the air was first detected on December 31st 2019 in China. Since then, with 3M confirmed cases and over 200K COVID-related deaths around the globe, the outbreak was declared a global pandemic on March 15th 2020 by the WHO. The rapid spread of the disease, the novelty of the virus combined with a lack of preparedness for such a global emergency has led to food, cleaning supply, and medicine shortages; delays in government-level decision making in matters such as the declaration of shelter-in-place orders and social distancing guidelines; shortages of life-saving medical devices such as ventilators and inadequate staffing at hospitals; overflowing morgues and bodies being stored in freezer trucks. Accurately estimating the number of future cases and deaths is essential for the optimal allocation of resources, the development of targeted and efficacious public health policies, and timely interventions and preparedness that will help curb further spread of the virus.

*First author's affiliation, 1156 High St, Santa Cruz, CA 95064

1.1 The Data

The data set consists of a 58 x 4 data frame containing information on the population, total number of COVID-19 cases, and the total number of deaths from COVID-19 for each of the 58 counties in California as of 4/13/2020. The first variable, 'County', is a factor variable and the remaining three variables are of type integer.

1.2 Exploratory Data Analysis

We start EDA by examining the data set for NAs, missing values, or irregularities. After confirming that the data set is complete, we order the data set by the three numerical variables to gain insight into the incidence and mortality rates of COVID-19 in the various counties. The highest number of deaths due to COVID-19 is in LA with 320 deaths, followed by Santa Clara (60 deaths) and Riverside (50 deaths). LA also has the highest number of infections (9,420 cases) and is the most populous county in CA with a population of 10,039,108 people. San Diego and Riverside follow LA in COVID-19 incidence with 1,847 and 1,751 cases respectively. San Diego and Orange county are the second and third most populous counties in CA with populations of 3,337,685 and 2,423,266 individuals. CA has a total population of over 39 million individuals and as of 4/13/2020, there have been 725 deaths due to COVID-19 and a total of 24,303 confirmed cases of the disease in CA. 5 counties have no confirmed cases of COVID-19 and 21 counties have had no deaths resulting from the disease.

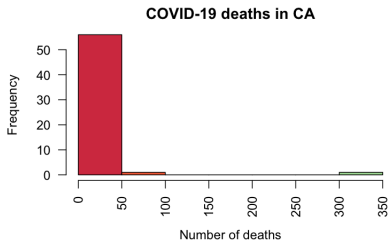

Figure 1: COVID-19 deaths per CA county.

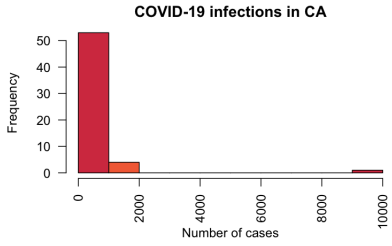

Figure 2: COVID-19 cases per CA county.

We then examine histograms for the number of deaths and cases for each county and find that the data are right-skewed with a potential outlier (Los Angeles).

We use a scatterplot to visualize the relationship between the number of cases and the number of deaths per county. There is a positive association between infections and deaths, but it is possible that this positive association is influenced by the large number of cases and deaths in LA.

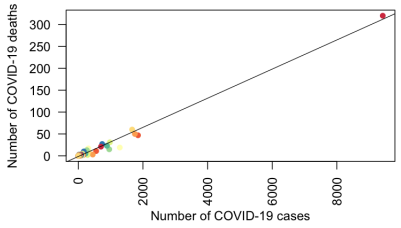

Figure 3: Number of COVID-19 cases vs deaths per CA county.

2. Methods

We explore three main models to make estimations of the number of deaths at the county level: a binomial model, a beta-binomial model that assumes the possibility of overdispersion of the data, and a hierarchical model. We utilize methods such as rejection sampling to sample from the posterior distributions and Laplace approximations, make transformations of variables (log, logit), and make use of tools such as contour plots to visualize outputs. We compare the three methods.

3. Analyses and Results

3.1 Model 1: Binomial

In the first model, we assume that the number of deaths follows a binomial distribution such that $y_i \sim Bin(n_i, \theta)$, We consider a beta prior on θ, the probability of death from COVID-19, such that $\theta \sim Be(1/2, 1/2)$.

Assuming independence between the counties, we calculate the posterior distribution of θ given y as:

$$p(\theta|y) = p(y|\theta)p(\theta)$$

$$\propto \prod_{i=1}^{58} \left[\theta^{y_i}(1-\theta)^{n_i-y_i}\right] \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$$

$$\propto \theta^{\sum_{i=1}^{58} y_i - \frac{1}{2}}(1-\theta)^{\sum_{i=1}^{58}(n_i-y_i)-\frac{1}{2}}$$

Thus, the posterior distribution of θ:

$$\theta|y \sim Be(\sum_{i=1}^{58} y_i + \frac{1}{2}, \sum_{i=1}^{58} n_i - \sum_{i=1}^{58} y_i + \frac{1}{2})$$

The expected value of $\theta|y$ can be found using the expression for the mean of the beta distribution, $E(X) = \frac{\alpha}{\alpha+\beta}$, as 0.02985105.

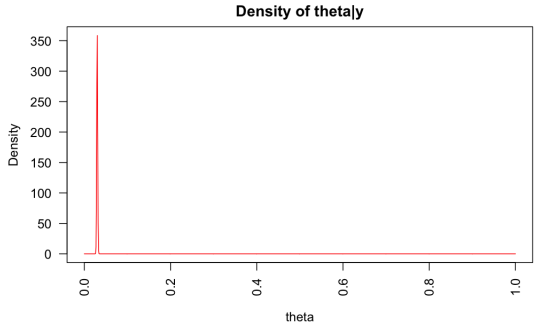

Figure 4: Posterior density of $\theta|y$ (n=1000)

Having calculated the posterior distribution, we can calculate the posterior predictive distribution to make predictions for a new county, $\tilde{y}$. We use the formula below:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)d\theta$$

$$= \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta$$

We will assume that 20% of the population will be infected. For simplicity, we will further assume that this will be the case at the county level i.e. 20% of the population in each county will be infected.

Assuming conditional independence between the number of deaths in current and future counties:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

with $p(\tilde{y}|\theta) = \binom{n}{\tilde{y}}\theta^{\tilde{y}}(1-\theta)^{n-\tilde{y}}$ and

$$\theta|y \sim Be(\sum_{i=1}^{58} y_i + \frac{1}{2}, \sum_{i=1}^{58} n_i - \sum_{i=1}^{58} y_i + \frac{1}{2})$$

For i=1:58, $p(\tilde{y}|y) =$

$$\frac{\binom{n}{\tilde{y}}B(\tilde{y}+\sum y_i+\frac{1}{2}, 2*\sum n_i - \sum y_i - \tilde{y} + \frac{1}{2})}{B(_i+\frac{1}{2}, \sum n_i \sum y_i + \frac{1}{2})}$$

3.1.1 Model 1 Predictions

In order to obtain future numbers of infection in CA, $\tilde{y}_1$, we draw samples from the posterior distribution of θ, plug in these samples to the binomial distribution for y and simulate 1000 draws. The lowest number of predicted deaths is 206,267 and the highest number of predicted deaths is 260,787. The average prediction of Model 1 is 235,693 deaths in CA.

As such, the assuming that 20% of the population of California becomes infected, the probability that 200,000 people will die of COVID-19 is 1.

3.2 Model 2: Beta-Binomial

We now consider a second model that assumes the possibility that the data are overdispersed i.e. the data show more variability than shown in the binomial model. We assume $y_i \sim BeBi(n_i, \mu, \tau)$ such that μ is the mean and τ is the precision parameter.

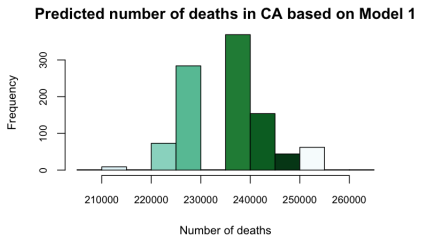

Figure 5: Model 1 predictions for number of COVID-19 deaths in CA

Then, the sampling distribution of y_i has density

$$p(y_i|\mu, \tau) = \binom{n_i}{y_i}\frac{B(\mu\tau + y_i, \tau(1-\mu) + n_i + y_i)}{B(\mu\tau, \tau(1-\mu))}$$

We consider a vague prior such that

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

We can compute the conditional posterior as below:

$$p(\mu, \tau|\mathbf{y}) = p(\mathbf{y}|\mu, \tau)p(\mu, \tau)$$

$$\propto \prod_{i=1}^{58} \frac{B(\mu\tau + y_i, \tau(1-\mu) + n_i + y_i)}{B(\mu\tau, \tau(1-\mu))} \times \frac{1}{\mu(1-\mu)(1+\tau)^2}$$

We do not have a closed form for the target distribution of $p(\mu, \tau|\mathbf{y})$. In order to achieve a sample from the posterior distribution, we can perform rejection sampling. As discussed in class and in the Jim Albert book, we will use a bivariate student t distribution as our proposal distribution q as this distribution is easy to sample from and has polynomially decaying tails that allow for more mass than the normal distribution. We will fix the location of the t at the mode,

which we will estimate using Laplace approximation, and we will use an inflated covariance matrix as the scale matrix. Then we find the bounding constant value c such that $p(\mu, \tau|y) \leq cq(\mu, \tau)$ for all μ and τ. In log form, this is the same as maximizing $logp(\mu, \tau|y) - logq(\mu, \tau)$ to find the constant d=log c such that $logp(\mu, \tau|y) - logq(\mu, \tau) \leq d$.

We begin with a reparameterization of μ and τ to the real line. We will log transform τ as it is positive valued, and logit transform μ as it is a proportion.

$$\theta_1 = logit(\mu) = log\left(\frac{1}{1-\mu}\right)$$

$$\theta_2 = log(\tau)$$

Then we can compute the posterior density of the transformed parameters by evaluating the density function at the inverse $\left(\frac{e^{\theta_1}}{1+e^{\theta_1}}, e^{\theta_2}\right)$ and multiplying by the Jacobian $\left(\frac{e^{\theta_1+\theta_2}}{(1+e^{\theta_1})^2}\right)$. Below is the contour plot of the transformed posterior:

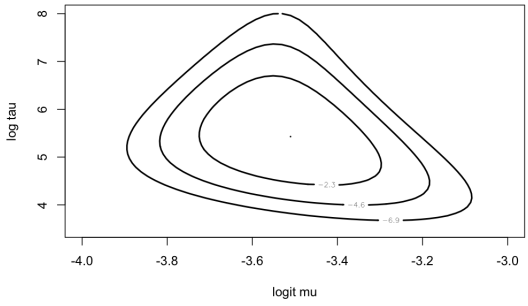

Figure 6: Contour plot of the log transformed value of the posterior

Based on this contour plot, we use $(-3.5, 5.4)$ as our starting point for the Laplace approximation. The Laplace approximation gives $(-3.518969, 5.455915)$ as the posterior mode, which is where the maximum value of d occurs. If we transform μ and τ back to their original parameterization, the posterior mode is $(0.0288, 234.139)$, with 0.0288 corresponding to the expected mortality rate and 234.139 corresponding to the standard error under the beta-binomial model.

We evaluate $logp(\mu, \tau|y) - logq(\mu, \tau)$ (log transformation of the ratio of the true distribution, the beta-binomial, and the proposal distribution, the bivariate t) at $(-3.518969, 5.455915)$ to get the maximum value of d, which is -3257.146.

Now that we have the value d at which the $logp(\mu, \tau|y) - logq(\mu, \tau)$ is maximized, we perform rejection sampling to sample from the target posterior distribution.

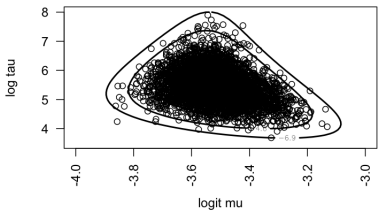

Figure 7: Contour plot with samples from rejection sampling.

We can see that most draws from the bivariate t distribution fall within the inner contour of the exact density. The acceptance rate is 51%.

3.2.1 Model 2 Predictions

In order to obtain future numbers of infection in CA under Model 2, $\tilde{y}_2$, we plug in the samples obtained by rejection sampling to the beta-binomial distribution for y and simulate 1000 draws. The lowest number of predicted deaths is 30,344 and the highest number of predicted deaths is 784,122. As discussed in class, the beta-binomial model allows for greater variation compared to the binomial model (Model 1). It is possible that Los Angeles was a particularly influential county for the estimates from the beta-binomial model, which could be confirmed via the leave-one-out analysis. The average prediction of Model 2 is 228,008 deaths in CA due to COVID-19.

As such, assuming that 20% of the population of California will become infected, the probability that 200,000 people will die of COVID-19 is 57.6%.

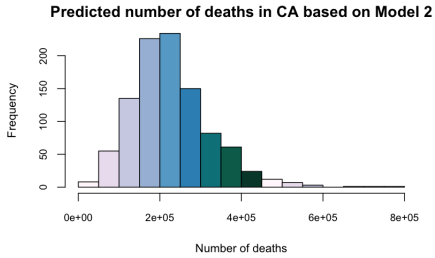

Figure 8: Model 2 predictions for number of COVID-19 deaths in CA

3.3 Model 3: Hierarchical

We build a third, hierarchical model which incorporates and quantifies the uncertainty associated with μ and τ. We consider the following:

$$y_i \sim Bin(n_i, \theta_i)$$
$$\theta_i \sim Be(\mu\tau, (1-\mu)\tau)$$
$$p(\mu, \tau) = (\mu(1-\mu)(1+\tau)^2)^{-1}$$

The joint posterior is

$$p(\boldsymbol{\theta}, \mu, \tau|\mathbf{y}) \quad \propto \quad p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

which we can factorize as

$$p(\boldsymbol{\theta}, \mu, \tau|\mathbf{y}) = p(\boldsymbol{\theta}|\mu, \tau, \mathbf{y})p(\mu, \tau|\mathbf{y})$$

We can obtain samples from the posterior distribution of $\boldsymbol{\theta}, \mu$ and τ via this factorization by sampling from $p(\mu, \tau|\mathbf{y})$, plugging these samples into $p(\boldsymbol{\theta}|\mu, \tau, \mathbf{y})$, and sampling from this posterior.

We first obtain an expression for $p(\theta_i|\mu, \tau, y_i)$:

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

Then

$$\theta_i|\mu, \tau, y_i \sim Beta(y_i+\mu\tau, n_i-y_i+\tau(1-\mu))$$

therefore $p(\boldsymbol{\theta}|\mu, \tau, \mathbf{y})$

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

and thus we can obtain samples from this block in the factorization using the closed form beta.

We next obtain an expression for $p(\mu, \tau|\mathbf{y})$ by integrating out θ from the joint posterior:

$$p(\mu, \tau|\mathbf{y}) = \int p(\mu, \tau, \boldsymbol{\theta}|\mathbf{y})d\theta$$
$$\propto p(\mu, \tau)\prod_{i=1}^{58}\frac{B(y_i+\mu\tau, n_i-y_i+\tau(1-\mu))}{B(\mu\tau, \tau(1-\mu))}$$

Then

$$p(\mu, \tau|\mathbf{y}) \propto (\mu(1-\mu)(1+\tau)^2)^{-1}\times$$
$$\prod_{i=1}^{58}\frac{B(y_i+\mu\tau, n_i-y_i+\tau(1-\mu))}{B(\mu\tau, \tau(1-\mu))}$$

We previously discussed the log and logit transformations and rejection sampling method to utilize while working with this distribution that does not have a closed form.

Now that we have the necessary blocks expressed, we can generate samples from the posterior distribution. Utilizing the rejection sampling methodology, we draw 10000 samples from $p(\mu, \tau|\mathbf{y})$, plug these samples into $p(\boldsymbol{\theta}|\mu, \tau, \mathbf{y})$ and draw 10000 from this resulting factor.

3.3.1 Model 3 Predictions

In order to obtain future numbers of infection in CA under Model 3, $\tilde{y}_3$, we plug in the samples obtained by rejection sampling to the beta distribution to simulate 1000 draws, and we draw 1000 $\tilde{y}_3$ from the predictive posterior distribution.

The lowest number of predicted deaths is 8,327 and the highest number of total predicted deaths in CA is 695,089. The average prediction of Model 3 is 229,638 total deaths in CA due to COVID-19.

As such, assuming that 20% of the population of California will become infected, the probability that 200,000 people will die of COVID-19 is 59.5%.

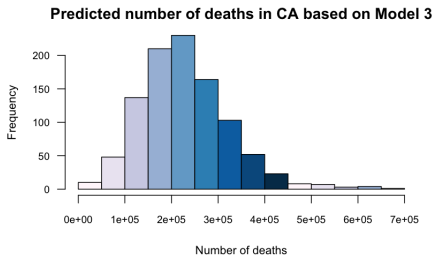

Figure 9: Model 3 predictions for number of COVID-19 deaths in CA

4. Conclusions

In this report we considered three models for predicting the number of COVID-19 deaths in the state of California.

We found that the binomial model, which predicted a 100% probability that over 200,000 people in CA will die due to COVID-19 under the assumption that 20% of the population will become infected, is overly simplistic for the data at hand as it does not account for the overdispersion in the data.

The next model, the beta-binomial, accounted for this problem by introducing two hyperparameters. Under the beta-binomial model, we predicted a 57.6% probability that over 200,000 people in CA will die due to COVID-19 under the assumption that 20% of the population will become infected. However, this model can be improved as it assumed that the hyperparameters are known, and therefore does not capture potential variability in the hyperparameters.

The third and final model we considered, the hierarchical model, solved this problem by considering priors for the hyperparameters.This model predicted a 59.5% probability that over 200,000 people in CA will die due to COVID-19 under the assumption that 20% of the population will become infected.

Taking into consideration the advantages and disadvantages of each model, we believe that the hierarchical model is the best model out of the three, closely followed by the beta-binomial model. The predictions from these two models agree.

While our focus in this paper was on prediction of deaths at the state-level, further investigations could explore predictions of deaths at county level. Results from such an exploration would allow for highly targeted policy making and preparation at the county level, and encourage higher collaboration and sharing of resources across counties.

In conclusion, the findings of this paper underscore the urgency and gravity of the COVID-19 pandemic within California, and highlight the importance of accurate and timely estimation of deaths during these unprecedented times.

REFERENCES

Albert, Jim. (2009) "Bayesian Computation with R". 2nd ed. Springer.

Harmon, Amy. 2020. 'Why We Don't Know the True Death Rate for Covid-19'. Retrieved April 24th, 2020 (https://www.nytimes.com/2020/04/17/us/coronavirus-death-rate.html).

World Health Organization (WHO). 2020. 'Coronavirus disease (COVID-19) Pandemic' (https://www.who.int/emergencies/diseases/novel-coronavirus-2019).