STA450 Assignment 1

Yu-Chun Chien

Question 1.

Introduction

As COVID-19 varies, the symptoms, transmission rate, and the recover rate differs. Observing COVID-19 data to date, many scientists hypothesized that the successive waves of COVID-19 are becoming narrower. It is also observed that the symptoms are becoming less severe, and the hospitalization rate are decreasing.

This section of the assignment aims to go through the data from Maharashtra state, New York, Belgium, and London and address the hypothesis that the successive waves of COVID-19 are becoming narrower. The concept of "narrower" will be conceptualized and assessed scientifically.

Method & Result

To address the question whether the waves are becoming narrower, the SIR model is utilized here to assess the waves in different cities. In SIR model, there are two parameters, β and γ, which represents the transmission rate and the recovery rate respectively. Dividing β by γ, we get R_0, which is the basic reproduction number. Under the SIR model, if $R_0 > 1$, the pandemic will take off, and if it < 1, the pandemic will die out. Also, the larger R_0 is, the "narrower" a wave is. Thus, in the following analysis, R_0 of different waves in the cities mentioned above will be calculated and compared. If R_0 is becoming bigger, then the hypothesis that the successive waves are narrower is correct. If it is not, then it might be that it is narrower only in some areas.

For the analysis, we set $S = 0.5$, $I = 0.001$, and $R = 1 - S - I$ for all cities and waves.

Maharashtra State, India

For Maharashtra state, observing the plot, it seemed that the successive wave is becoming narrower. According to the plot, the first wave occurred from 2020/08/15 to 2020/11/15, the second (Delta) wave occurred from 2021/03/15 to 2021/06/15, and the third (Omicron) wave occurred from 2021/12/30 to 2022/02/06. SIR model is fitted and the β, γ, and R_0 is showed in the table below.

Real data vs. Fitted SIR Model

The real data (points) and the fitted SIR model (line) is plotted below for the three waves. The model is fitted by minimizing the mean squared error. As you can see from the plot, the model fits well for the first and second wave, while it does not fit well in the third wave, which might be because that the data size of it is smaller.

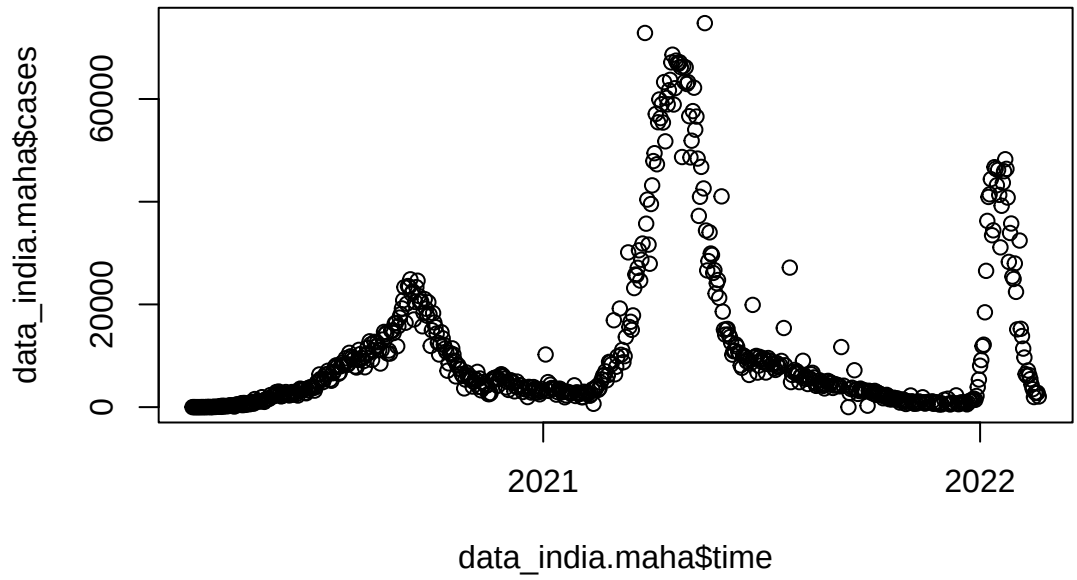
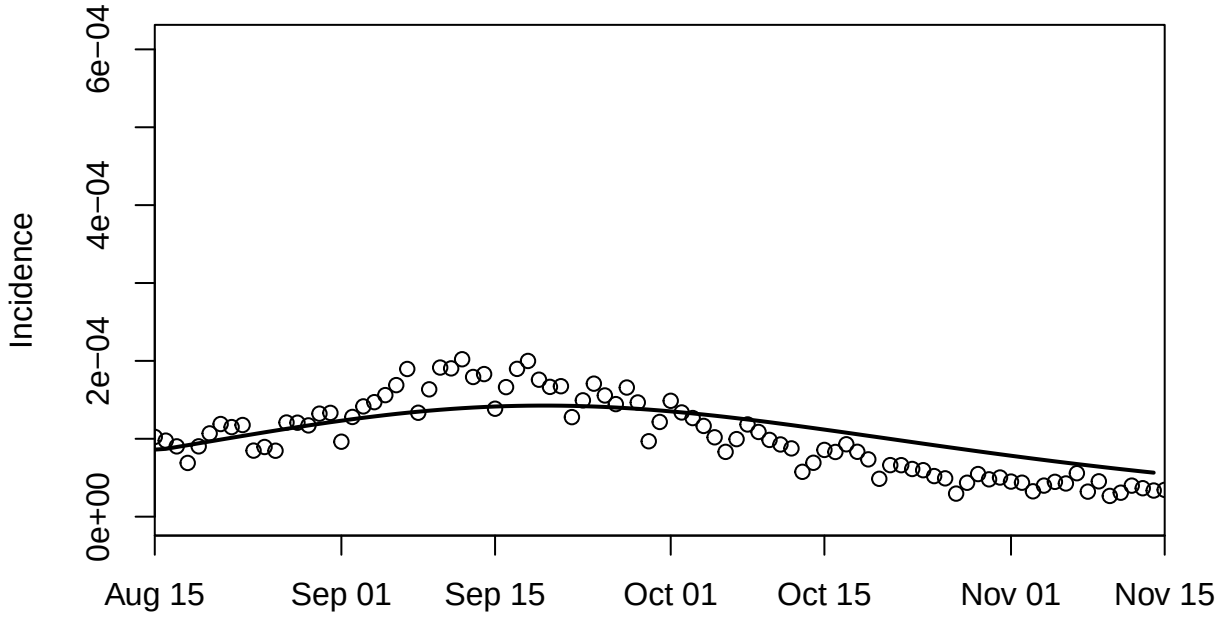
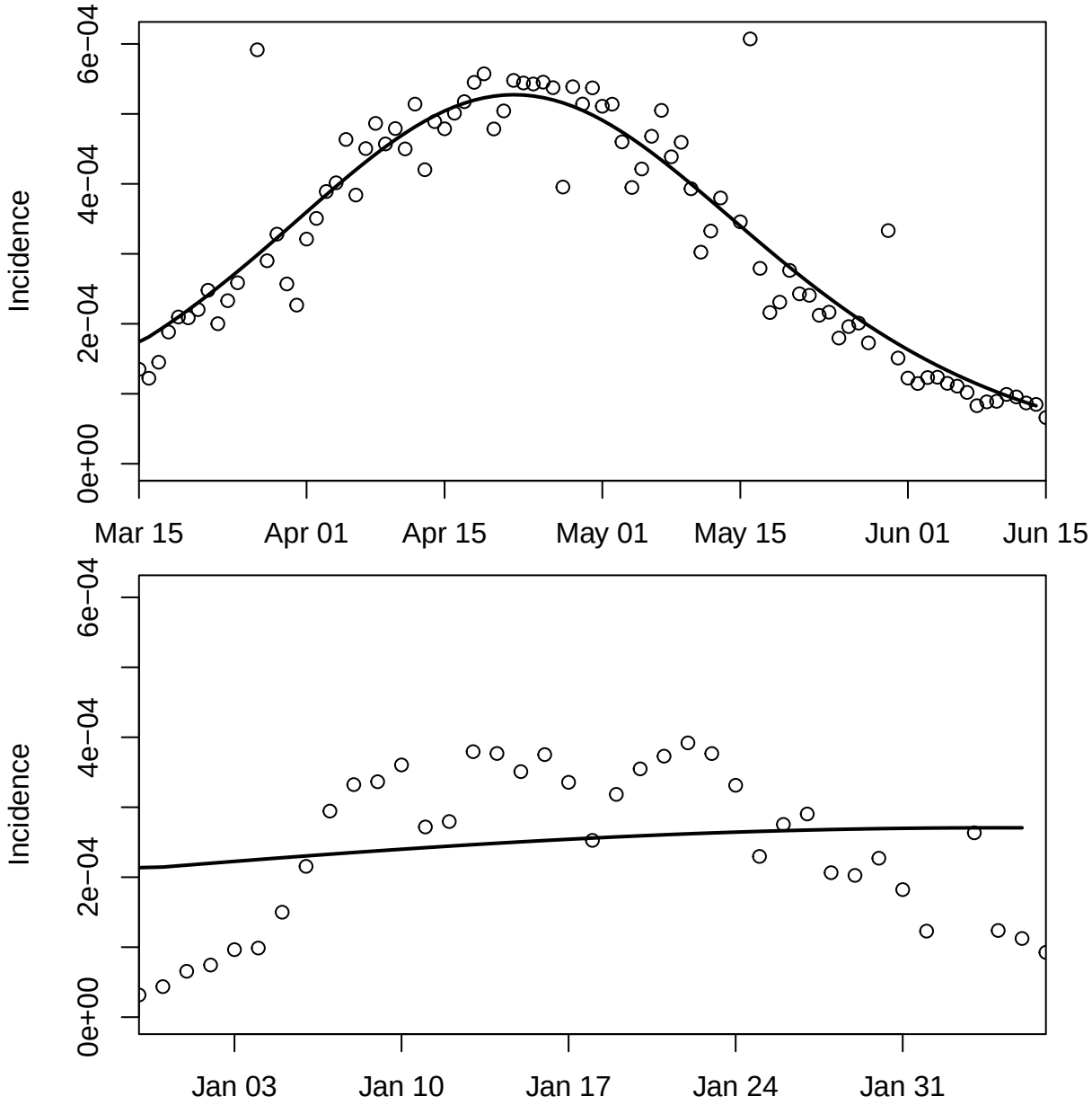


Figure 1: Cases in Mahasthra state



wave	beta	gamma	R_0
First Wave	4.98	0.05	99.600000
Second Wave	3.44	0.99	3.474747
Third Wave	1.46	0.50	2.920000



β, γ & R_0 of the Three Waves

By comparing the R_0 value of each wave, since the R_0 value is not increasing, it means that under the deterministic SIR model, our wave is not becoming narrower. However, this is contrary to the plot, as we could observe that the waves are becoming narrower. Thus, it might be that the SIR model is not suitable for modeling COVID-19, or simply minimizing the mean square error for observed vs. fitted data does not gives us the best parameter. In particular, it is commonly believed that the transmission rate is becomming higher while the transmission rate is lower.

New York, United States

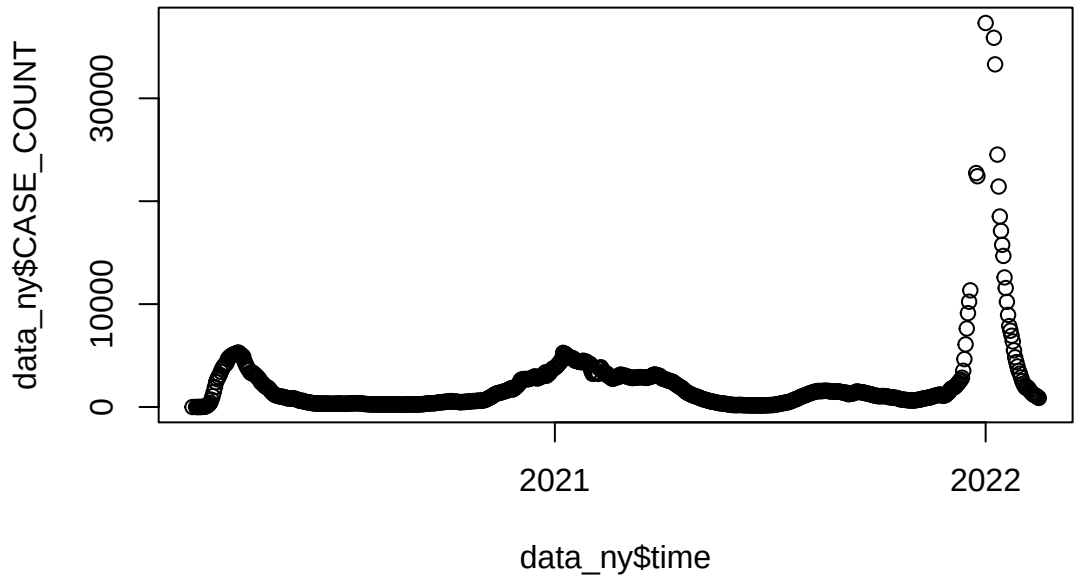
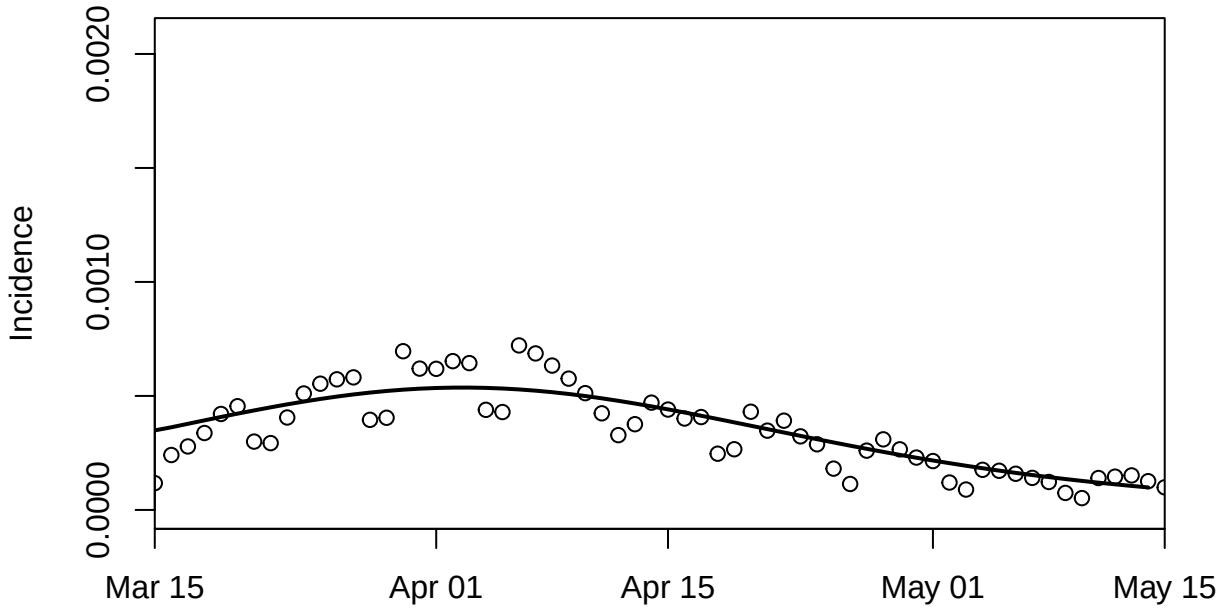


Figure 2: Cases in New York

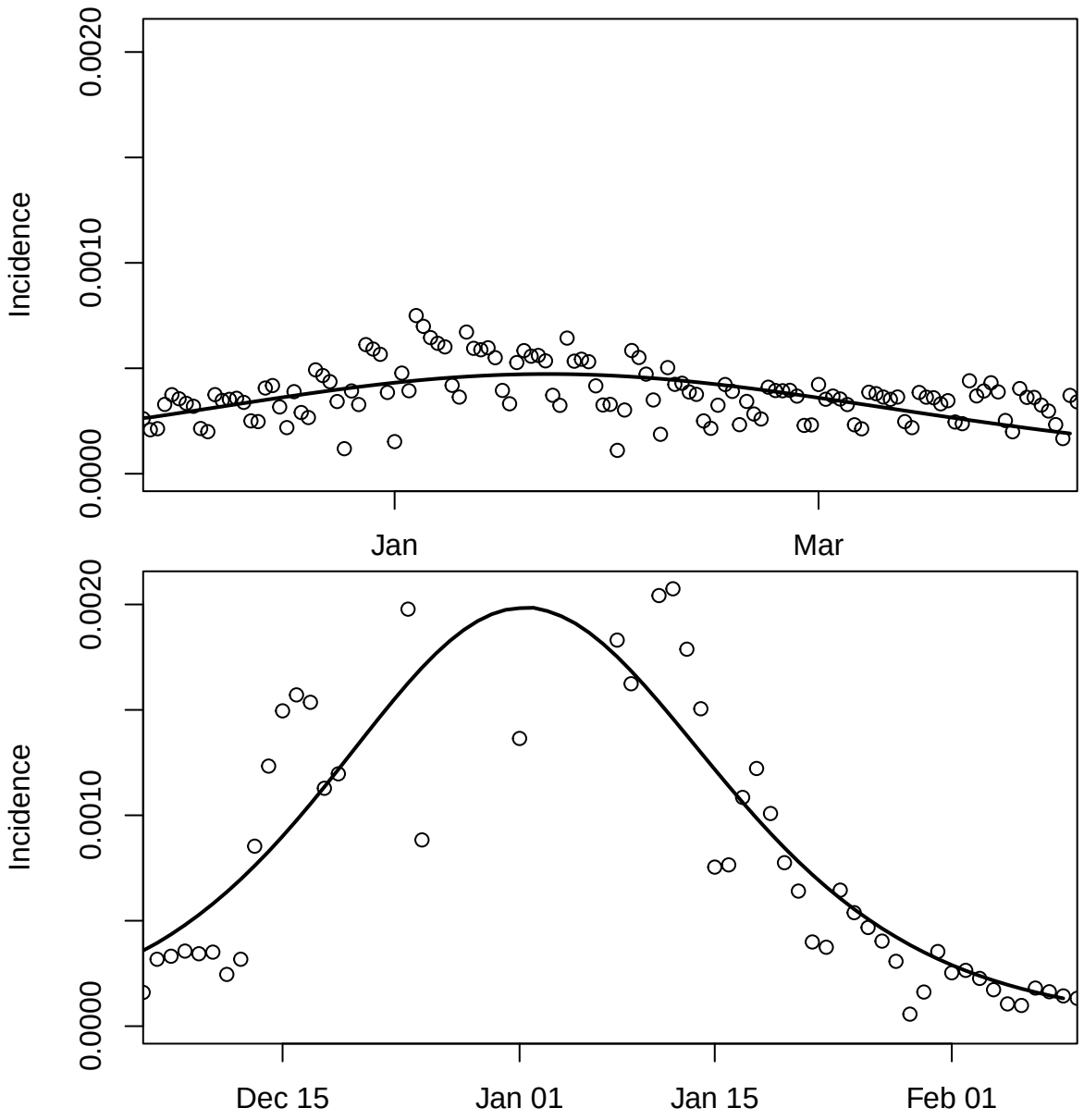
For New York, observing the plot, it seemed that the successive wave is not becoming narrower. The second (Delta) wave is the widest, while the third (Omicron) wave is the narrowest. According to the plot, the first wave occurred from 2020/03/15 to 2020/15/15, the second (Delta) wave occurred from 2020/11/27 to 2021/04/06, and the third (Omicron) wave occurred from 2021/12/05 to 2022/02/10. SIR model is fitted and the β, γ, and R_0 is showed in the table below.

Real data vs. Fitted SIR Model

The real data (points) and the fitted SIR model (line) is plotted below for the three waves. The model is fitted by minimizing the mean squared error. As you can see from the plot, the model fits well for all three waves.



wave	beta	gamma	R_0
First Wave	4.95	0.99	5.000000
Second Wave	0.86	0.54	1.592593
Third Wave	2.92	0.99	2.949495



β, γ & R_0

By comparing the R_0 value of each wave, the second wave has the smallest value while the first wave have the largest. However, the result does not correspond to the plot. To match our predicted results observed from the plot, the third wave should have the highest R_0 value. Thus, it might be that the SIR model is not suitable for modeling COVID-19, or simply minimizing the mean square error for observed vs. fitted data does not gives us the best parameter. In particular, it is commonly believed that the transmission rate is becoming higher while the transmission rate is lower.

Belgium

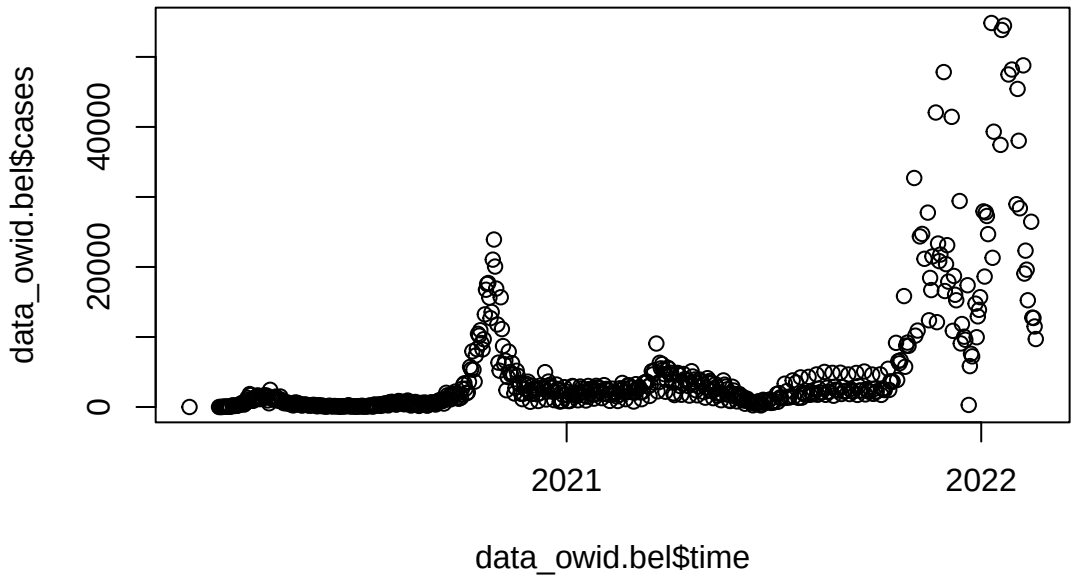


Figure 3: Cases in Belgium

For Belgium, observing the plot, it seemed that the successive wave is becoming narrower. Also, there seemed to be only two waves instead of three. According to the plot, the first wave occurred from 2020/10/15 to 2020/12/15, and the second (Omicron) wave occurred from 2021/10/15 to 2022/02/01. SIR model is fitted and the β, γ, and R_0 is showed in the table below.

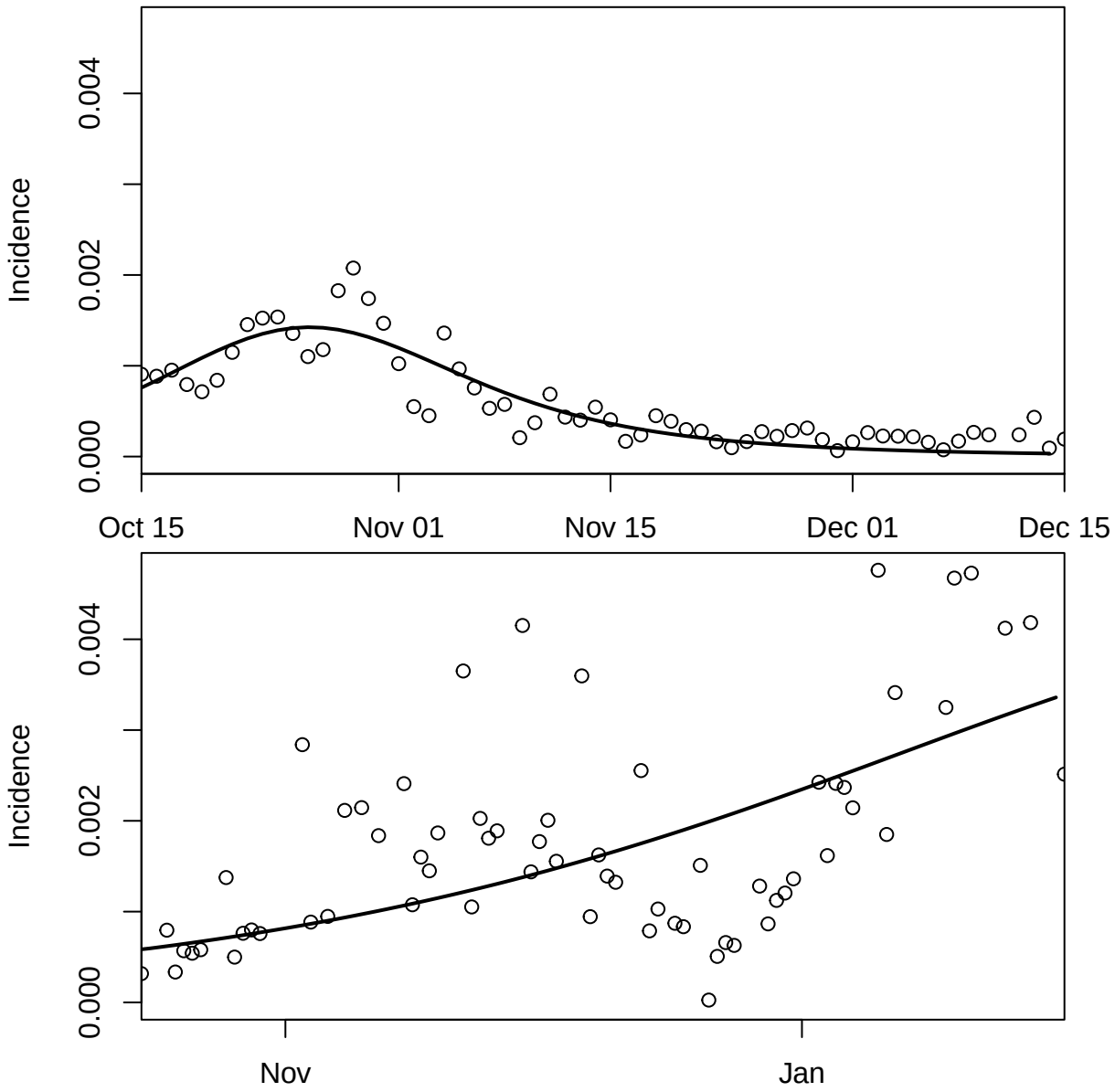
Real data vs. Fitted SIR Model

The real data (points) and the fitted SIR model (line) is plotted below for the three waves. The model is fitted by minimizing the mean squared error. As you can see from the plot, the model fits well for both the first and the second wave.

β, γ & R_0

By comparing the R_0 value of each wave, the value of R_0 for the second wave is smaller than the first wave, which implies that the first way is narrower. This is a contradiction to the plot. It might be that the SIR model is not suitable for modeling COVID-19, or simply minimizing the mean square error for observed vs. fitted data does not gives us the best parameter. In particular, it is commonly believed that the transmission rate is becoming higher while the transmission rate is lower.

wave	beta	gamma	R_0
First Wave	5.12	0.060	85.33333
Second Wave	0.03	0.004	7.50000



London, UK

For London, observing the plot, the successive waves do not seem to vary a lot, although the height is different. According to the plot, the first wave occurred from 2020/10/15 to 2020/02/15, the second (delta) wave occurred from 2021/06/15 to 2021/10/15, and the third wave occured from 2021/12/08 to 2022/02/08. SIR model is fitted and the β, γ, and R_0 is showed in the table below.

Real data vs. Fitted SIR Model

The real data (points) and the fitted SIR model (line) is plotted below for the three waves. The model is fitted by minimizing the mean squared error. As you can see from the plot, the model fits well for all three

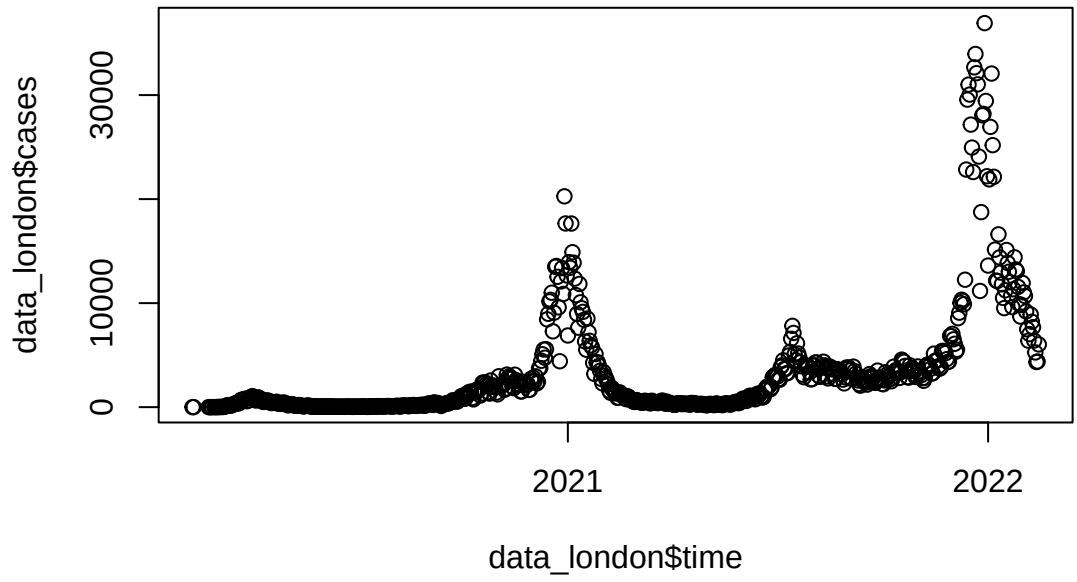
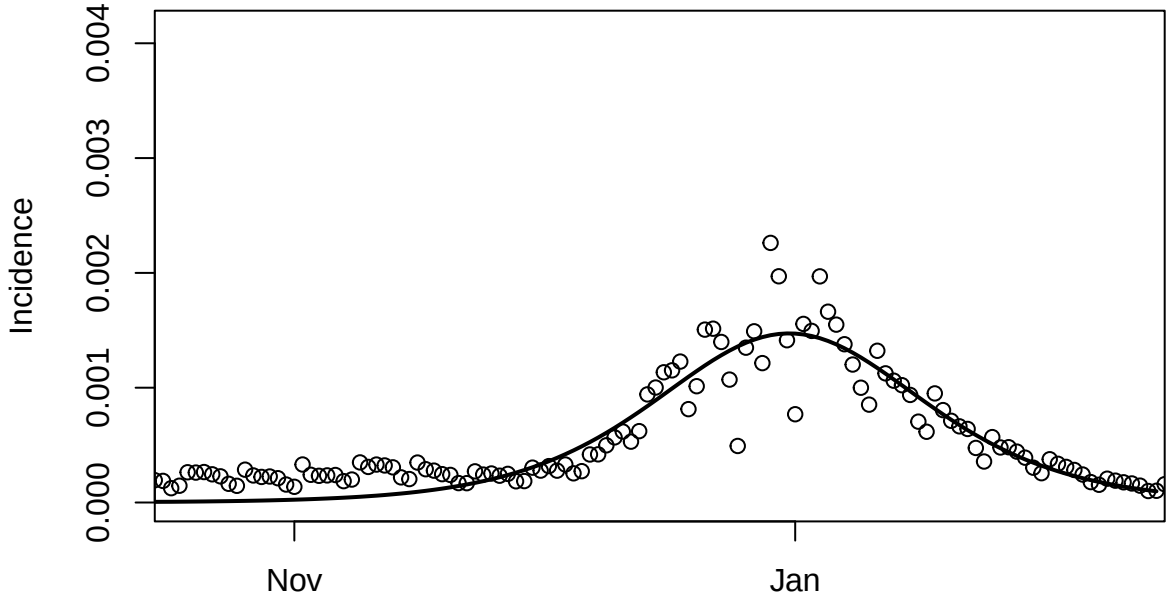
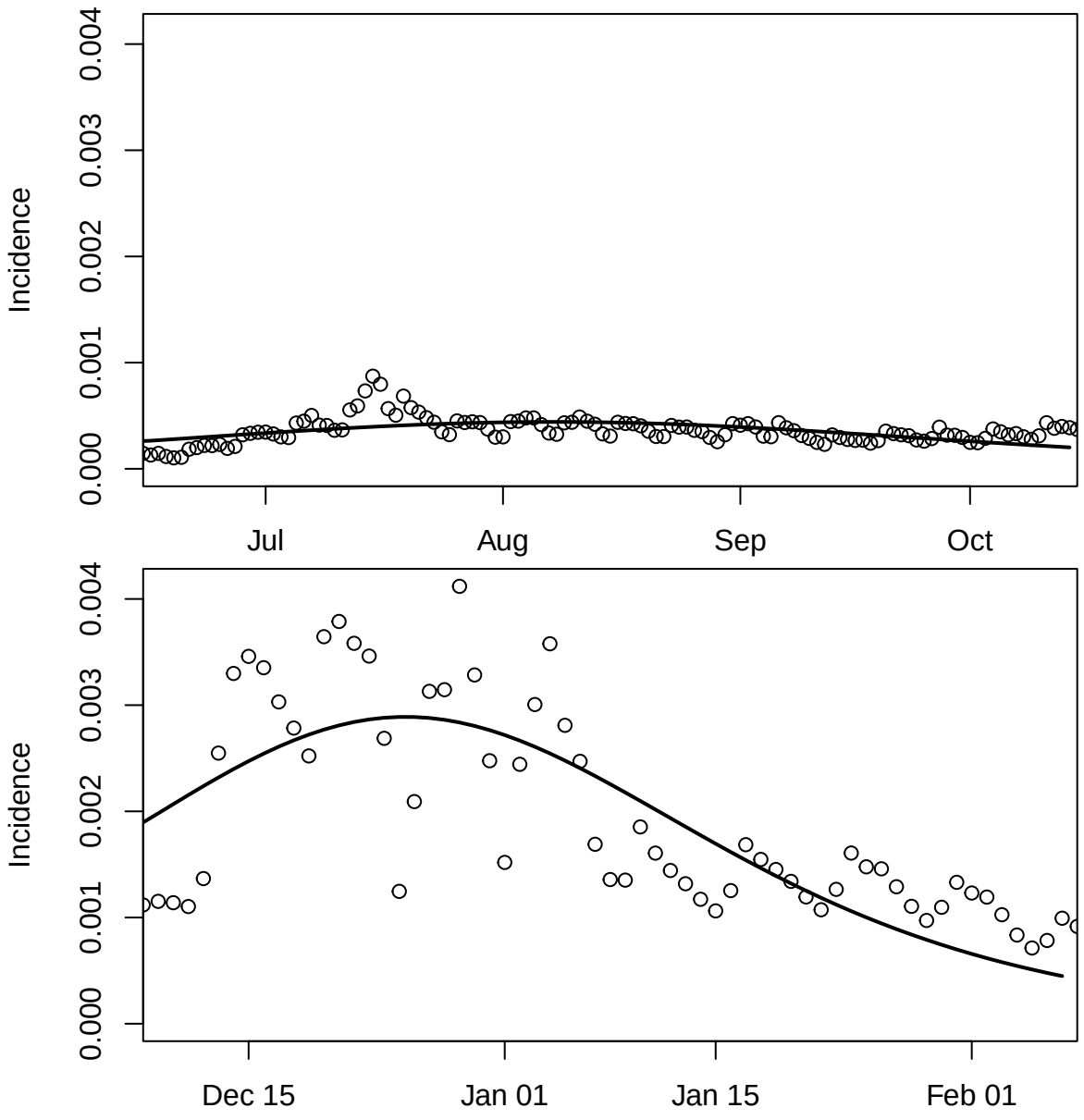


Figure 4: Cases in London

waves.



wave	beta	gamma	R_0
First Wave	1.41	0.01	141.000000
Second Wave	0.90	0.53	1.698113
Third Wave	1.00	0.36	2.777778



β, γ & R_0

By comparing the R_0 value of each wave, the value of R_0 for the first wave is way bigger than the second and third wave, while according to our plot it is not true since the shape of the three waves in London are approximately the same except for the height. It might be that the SIR model is not suitable for modeling COVID-19, or simply minimizing the mean square error for observed vs. fitted data does not gives us the best parameter. In particular, it is commonly believed that the transmission rate is becoming higher while the transmission rate is lower.

Conclusion

By fitting the SIR Model and comparing R_0 across different waves, it could be observed that the value is not consistently becoming larger, and some β and γ is not making sense when comparing to our common findings that the variant are becoming more infectious but have a faster recovery rate. Thus, the hypothesis that the successive waves are narrowing is not correct if we apply the SIR model and fit the model to the data by minimizing the mean squared error. However, it might be that globally, it is narrowing, or we should apply another model and use another optimization method.

Question 2

Introduction

When people got COVID-19, it often takes time for them to go take the test and be notified of the infection. It is hypothesized that this kind of reporting delays make SIR models harder to identify. If it takes time from infection and notification of the infection, a SINR model is used to model the disease, where N stands for notification.

To understand how reporting delays affect the SIR modeling, this section aims to simulate a stochastic epidemic with 100 individuals using the parameters $\beta = 3.441975$ and $\gamma = 0.9910306$ estimated from the Delta wave in Maharashtra in Question 1. Three scenarios are presented, one with no reporting delay, one with 2 days of delay on average, and one with 5 days of delay on average. If the hypothesis is true, it is predicted that the longer the reporting delay is, the "wider" the posterior distribution of the parameters will be. Namely, the posterior distribution of the parameter of the 5 day delay scenario will be widest, with larger variance.

Methods & Results

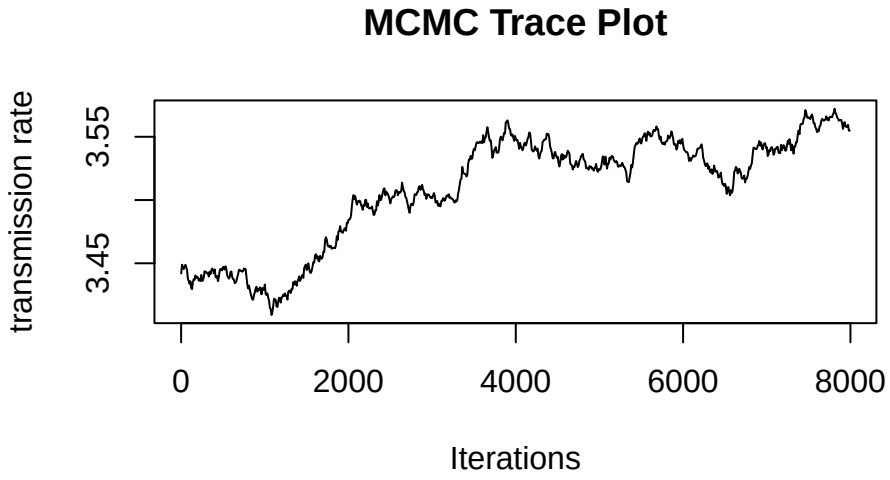
To simulate a stochastic epidemic, Markov chain Monte Carlo (MCMC) is used. In the no delay scenario, 8000 iterations are run with each having 1 chains. In the 2 days and 5 days delay scenario, 500 iterations are run with each having 1 chains. In this section, three scenario will be discussed separately.

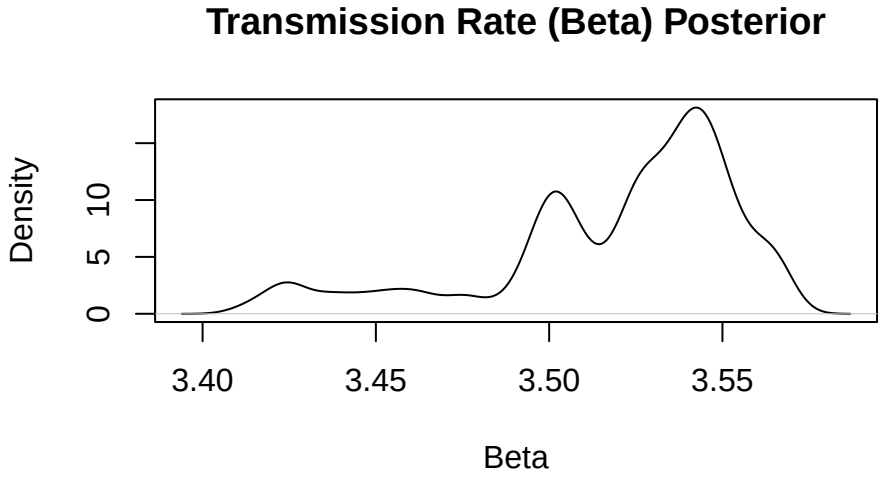
No Delay

For the no delay condition, using a $\gamma(1, 0.99)$ prior for infectious periods, a $\gamma(20, \frac{20}{3.441975})$ prior on beta, and using 0.001 for the proposal variance for the MCMC update, 5000 simulations were simulated. Here, the shape parameter for the infectious period were chosen to be 1 since we assume that the distribution between two events follows a exponential distribution, and the rate parameter is 0.99, which is the recover rate. The prior on beta is chosen to be not too narrow to let our data to have more influence on the posterior.

MCMC Trace Plot

```
## ************************************************
## Start performing MCMC for the  known epidemic  SIR ILM for
## 8000 iterations
## ************************************************
```





Posterior of Transmission Rate

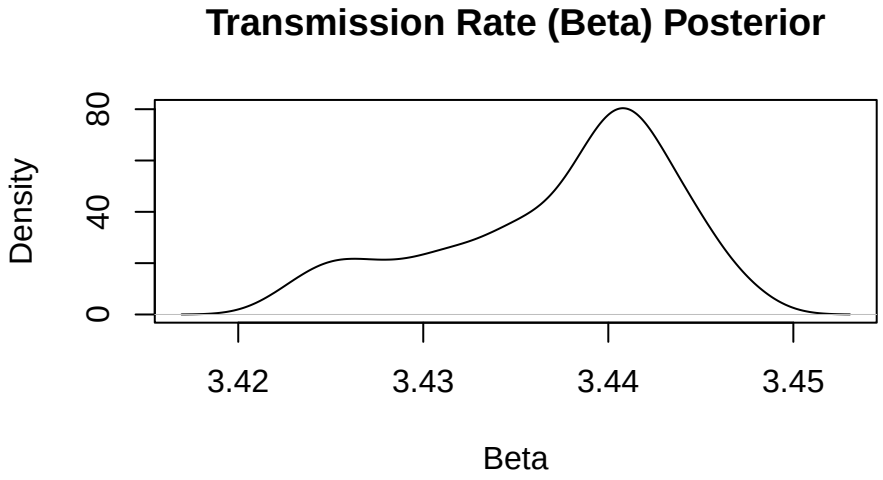
Here, due to the sample size and a lower iterations simulated, our transmission rate did not converge strongly. However, in general, it tends to converge to around 3.55.

```
##
## Iterations = 1000:8000
## Thinning interval = 10
## Number of chains = 1
## Sample size per chain = 701
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       3.518856       0.036575       0.001381       0.020532
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
## 3.424 3.502 3.529 3.545 3.566
```

For the posterior of the parameter, the mean is centered at 3.519, with the standard deviation being 0.037 and its 95 confidence interval being $(3.424, 3.566)$. Also, our acceptance rate is higher than 60% here.

2 Days Delay

For 2 days delay period condition, the incubation rate is 0.5 ($\frac{1}{days}$). The prior distribution of incubation periods is set to be $\gamma(1, 0.5)$, which is $\exp \gamma^{inc}$. The prior of delay period is set to be $\gamma(1, 0.5)$, which is $\exp \gamma^{del}$. Further, we force spark parameter to be low, so we set the initial value low and give it a narrow prior.

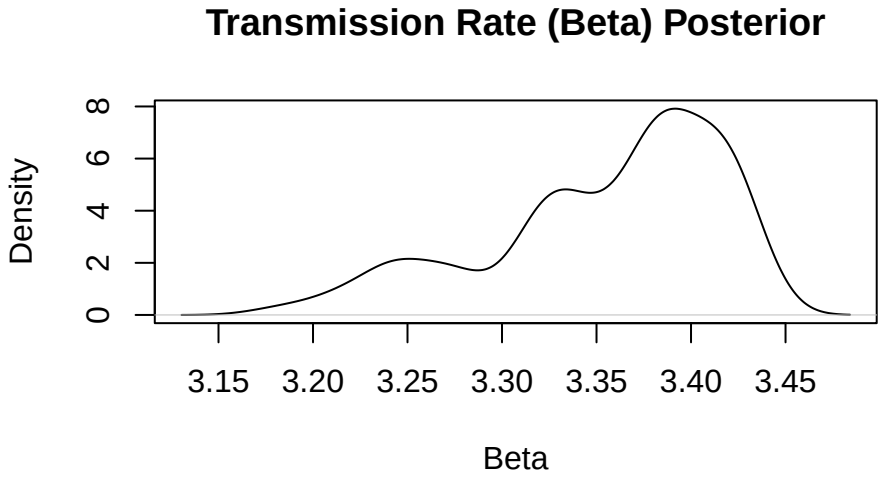



```
##
## Iterations = 100:500
## Thinning interval = 100
## Number of chains = 1
## Sample size per chain = 5
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                             Mean        SD  Naive SE Time-series SE
## Alpha_s[1]             3.439e+00 7.335e-03 3.280e-03      3.280e-03
## Spark                  1.096e-05 5.582e-06 2.496e-06      2.496e-06
## Incubation period rate 6.034e-01 4.872e-02 2.179e-02      2.179e-02
## Delay period rate      5.669e-01 1.074e-01 4.803e-02      4.803e-02
##
## 2. Quantiles for each variable:
##
##                             2.5%       25%       50%       75%     97.5%
## Alpha_s[1]             3.430e+00 3.436e+00 3.439e+00 3.444e+00 3.448e+00
## Spark                  4.427e-06 5.485e-06 1.411e-05 1.519e-05 1.566e-05
## Incubation period rate 5.433e-01 5.970e-01 6.007e-01 6.073e-01 6.678e-01
## Delay period rate      4.079e-01 5.527e-01 5.911e-01 6.259e-01 6.681e-01
```

For the posterior of the parameter, the mean is centered at 3.442, with the standard deviation being 0.009 and its 95 confidence interval being $(3.43, 3.45)$. Also, our acceptance rate is higher than 60% here.

5 Days Delay

For 5 days delay period condition, the incubation rate is 0.2 ($\frac{1}{days}$). The prior distribution of incubation periods is set to be $\gamma(1, 0.2)$, which is $\exp \gamma^{inc}$. The prior of delay period is set to be $\gamma(1, 0.8)$, which is $\exp \gamma^{del}$. Further, we force spark parameter to be low, so we set the initial value low and give it a narrow prior.




```
##
## Iterations = 100:500
## Thinning interval = 100
## Number of chains = 1
## Sample size per chain = 5
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                             Mean        SD  Naive SE Time-series SE
## Alpha_s[1]             3.338e+00 9.490e-02 4.244e-02      4.244e-02
## Spark                  2.992e-06 2.711e-06 1.213e-06      1.213e-06
## Incubation period rate 2.113e-01 1.394e-02 6.233e-03      6.233e-03
## Delay period rate      8.018e-01 6.440e-02 2.880e-02      2.880e-02
##
## 2. Quantiles for each variable:
##
##                             2.5%       25%       50%       75%     97.5%
## Alpha_s[1]             3.194e+00 3.330e+00 3.371e+00 3.387e+00 3.419e+00
## Spark                  8.006e-07 1.098e-06 1.204e-06 5.609e-06 6.215e-06
## Incubation period rate 1.939e-01 2.072e-01 2.081e-01 2.189e-01 2.286e-01
## Delay period rate      7.374e-01 7.765e-01 7.805e-01 8.141e-01 8.958e-01
```

For the posterior of the parameter, the mean is centered at 3.358, with the standard deviation being 0.098 and its 95 confidence interval being $(3.22, 3.45)$. Also, our acceptance rate is higher than 60% here.

Discussion, Interpretation, & Limitations

As you can see from the results and plots, the longer the delay period is, the wider the posterior distributions are. This means that the longer the delay period is, the more uncertain we are for the true value and distribution of the parameters. In the case of COVID-19, in the middle of a wave, we might expect a longer reporting time since many tests needs to be done or there might not be enough testing kits, making it harder to model the disease using SIR model since we might not be too certain about the parameter

Furthermore, MCMC in general takes longer time to run, especially when we have incubation and the delay period. Thus, only 100 individuals were used for sample size. In addition, for 2 days and 5 days delay, only 500 iterations were simulated. For no delay scenario, the run time is shorter, so 5000 iterations were simulated. This might be a potential limitation. To make MCMC simulations work better or to better estimate the parameters, more iterations should be run and a larger sample size should be used by using a computer that has better computing capacity.