Camilo Velez R - James Boggs

Midterm Project

Radu Balan

University of Maryland - Spring 2023

Introduction

For this project, we compared COVID-19 infections and deaths data from Manhattan (NYC) and Washington DC. For each, we establish models using SIR and SEIR epidemiologic models. For our computations, we used Julia, which was an incredible tool given that these were heavy computations and we were able to make parts of our code thread-safe and implement multi-threading.

Overall, it was a great experience to extract valuable information from these complex models. We gained a lot of experience in scientific computation with Julia and it was rewarding to see the models studied in class applied to actual data.

Background/Methods

The models that we used for this problem were the SIR and SEIR epidemiologic models. These models divide a population into groups (compartments), the **S**usceptible group which hasn't caught the disease yet, the **I**nfected group who currently has the disease and is contagious, and the **R**ecovered group who had the disease but no longer does. SEIR adds one more category for **E**xposed people who have the disease but are not yet contagious. These groups are chosen based on our understanding of how infectious diseases spread. Movement of people between these groups is described by the following differential equations.

- SIR Model:

  $\frac{dS}{dt} = -\beta S\frac{I}{N},\ S(0) = N$

  $\frac{dI}{dT} = \beta S\frac{I}{S} - \alpha I$

  $\frac{dR}{dT} = \alpha I,\ R(0) = 0$

- SEIR Model:

  $\frac{dS}{dt} = -\beta S\frac{I}{N},\ S(0) = N$
  $\frac{dE}{dt} = \beta S\frac{I}{S} - \delta E \frac{dI}{dt} = \delta E - \alpha I,\ I(0)\ \frac{dR}{dt} = \alpha I,\ R(0) = 0$

$$\frac{dI}{dt} = \delta E - \alpha I$$

$$\frac{dR}{dt} = \alpha I, \; R(0) = 0$$

The differential equations introduce the parameters α, β, δ which describe how quickly people move between the categories and also N which is the total population. The description of each parameter are as follows:

- α: rate at which people move out of the infected category, 1/α is the average infectious period
- β: rate at which people move out of the susceptible category, β is the number of close contacts per day per one infected individual
- δ: rate at which people move out of the exposed category, 1/ δ is the average incubation period
- N: total population

When fitting our model we introduced some additional parameters:

- R_0: instead of trying different values of β directly we introduce R_0 = β/α, so with the α and R_0, β can be computed. Then R_0 can be interpreted as the average number of close contacts per infected individual
- $E(0)$: the number of people in the exposed category at t = 0
- $I(0)$: the number of people in the exposed category at t = 0
- γ: the death rate of the disease

This γ parameter was used to compute pseudo-category for deaths (Y) with the following equation:

$$Y(t) = \gamma R(t)$$

To tune our model so it matched the observed data well we used the objective functions to determine the most optimal set of parameters:

SIR:

$$J(\alpha, \beta, N, \gamma) = \sum_{t=0}^{T_{max}} \left| I(t) - I_{sim}(t) \right|^p + \lambda \sum_{t=0}^{T_{max}} \left| Y(t) - \gamma R_{sim}(t) \right|^p$$

For $1 \leq p < \infty$ and

$$J(\alpha,\ \beta,\ N, \gamma) = \left|I(t) - I_{sim}(t)\right| \ + \ \left|Y(t) - \gamma R_{sim}(t)\right|$$

For p = ∞

SEIR:

$$J(\alpha,\ \beta,\ \delta, N, \gamma) = \sum_{t=0}^{T_{max}} \left|I(t) - I_{sim}(t)\right|^p + \lambda \sum_{t=0}^{T_{max}} \left|Y(t) - \gamma R_{sim}(t)\right|^p$$

For $1 \leq p < \infty$ and

$$J(\alpha,\ \beta, \delta,\ N, \gamma) = \left|I(t) - I_{sim}(t)\right| \ + \ \left|Y(t) - \gamma R_{sim}(t)\right|$$

For p = ∞

And λ = 1

Then we used the following algorithm to search through our ranges of parameters and find the one that minimizes the objective function J.

- Detect the onset of infections (t_0) which is the first value of t where $V(T) \geq V_{min} = 5$
- From the cumulative infections data (V) compute active infections (I) with the equation $I(T) = V\left(t \ + \ t_0 + \ \tau_0\right) - V\left(t \ + \ t_0 - \ \tau_0\right),\ 0 \leq t \leq T_{max}$ where $\tau_0 = 7$
- For each combination of parameters
    - simulate the model using the Euler method to get $S_{sim},\ I_{sim},\ R_{sim}$ or $S_{sim},\ E_{sim,}\ I_{sim},\ R_{sim}$
    - Compute $\hat{\gamma} = argmin_{\gamma} \| Y \ - \ \gamma R_{sim} \|_p$
    - Compute the objective function $J(\alpha,\ \beta, N, \gamma)$ or $J(\alpha,\ \beta,\ \delta, N, \gamma)$
- Determine the minimum value of J and the set of parameters that minimize it

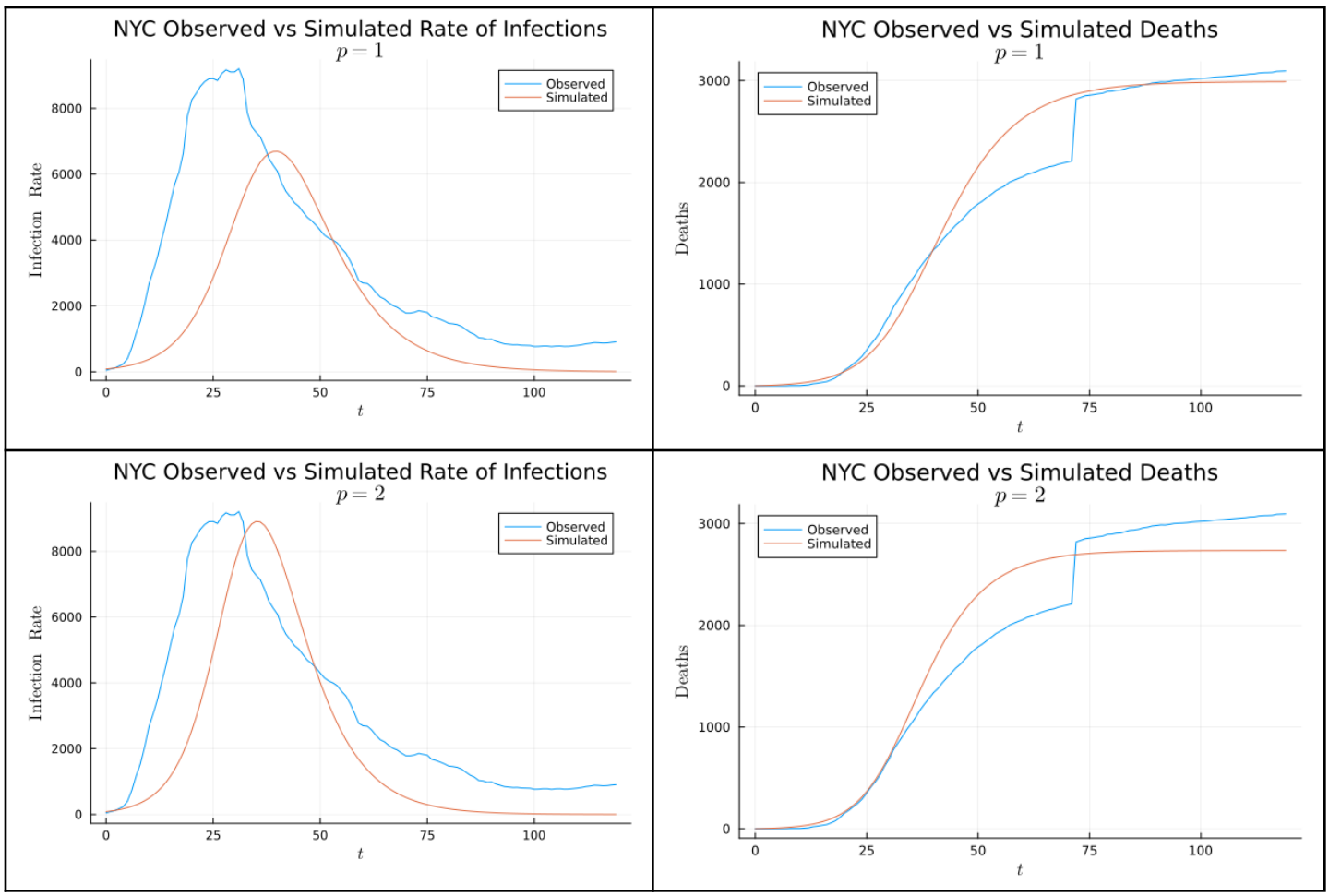
The algorithm takes in a large set of possible parameters and finds the most optimal set of parameters as measured by the objective function J

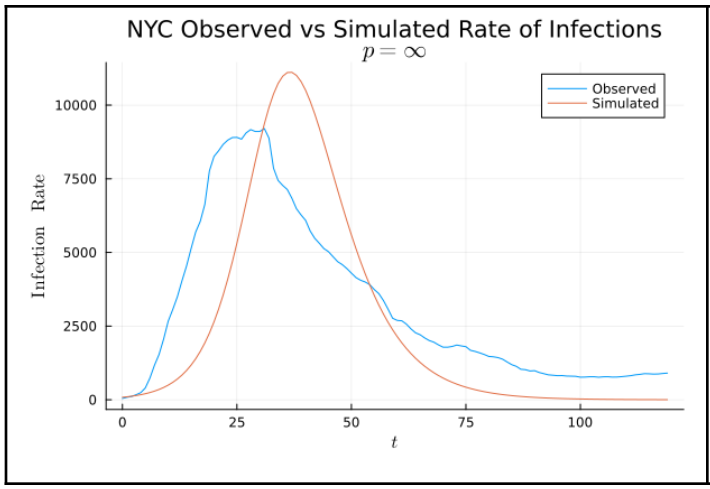
Results

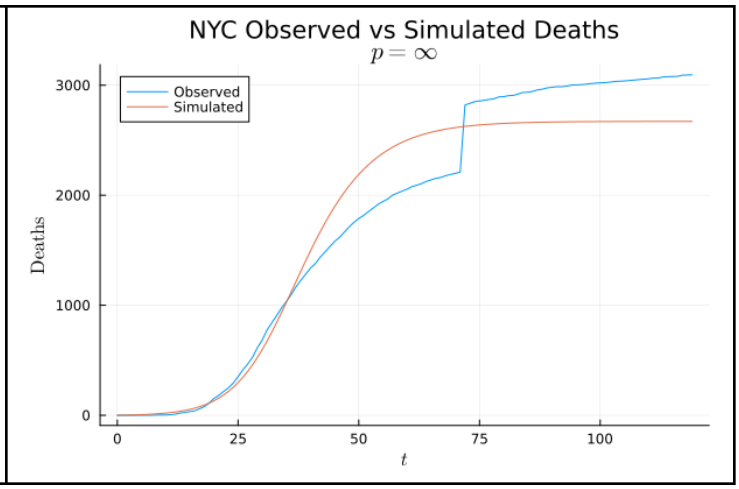
For the SIR model we searched the following range of parameters:

- α ∈ [0.05, 0.4]
- R_0 ∈ [1.5, 1.9]
- N / N_{max} ∈ [2%, 5%]

Here are the results for the SIR model fit to the first 120 days of the data on the Manhattan dataset ($T_{max} = 119$):







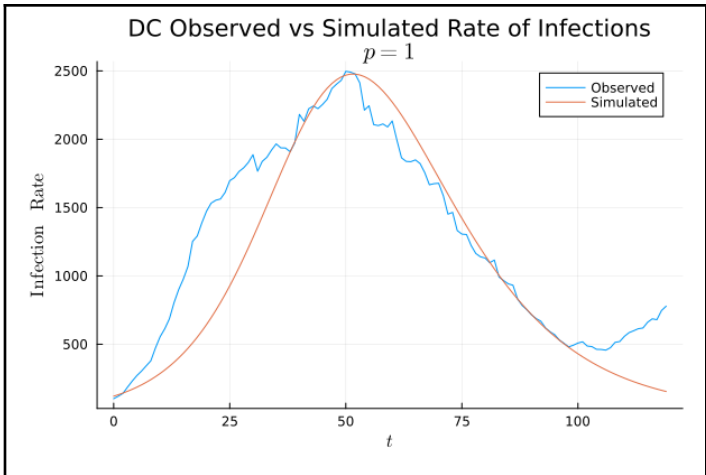
Parameters:

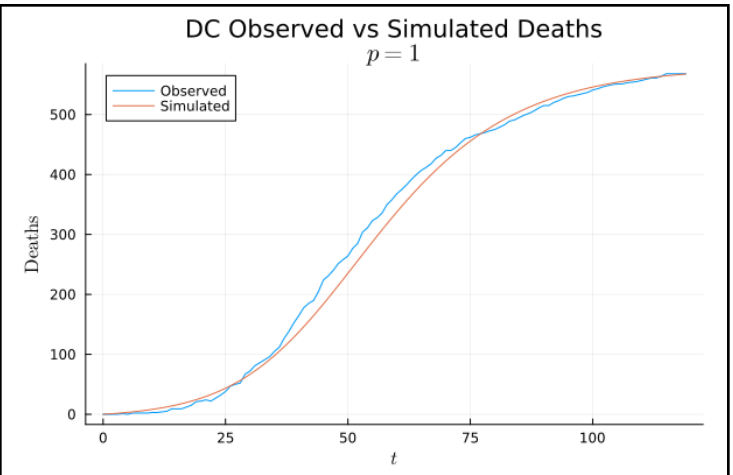
	p = 1	**P = 2**	**P = ∞**
α	0.17	0.2	0.2
β	0.323	0.38	0.38
R_0	1.9	1.9	1.9
N / N_{max}	3%	4%	5%
γ	0.08	0.05	0.043
J	209,877	5.33×10^8	6442

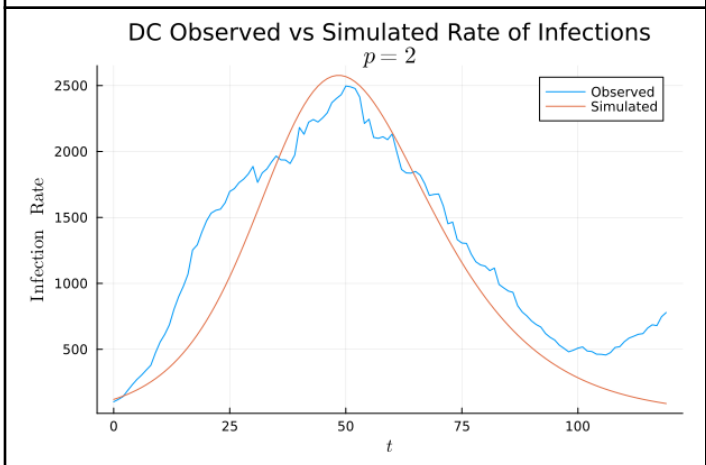
It was very interesting to see a gamma of 0.08 in p=1 since this was estimated to be the death rate of COVID-19 in NYC during this period by CDC data.[1] This is also evident in the graph of the estimated deaths for p=1: it is the most accurate of the three. All of our models failed to predict the timing of the first peak in infections, this can indicate that there were indeed many many cases of COVID during the first increase happening faster than our models could predict.

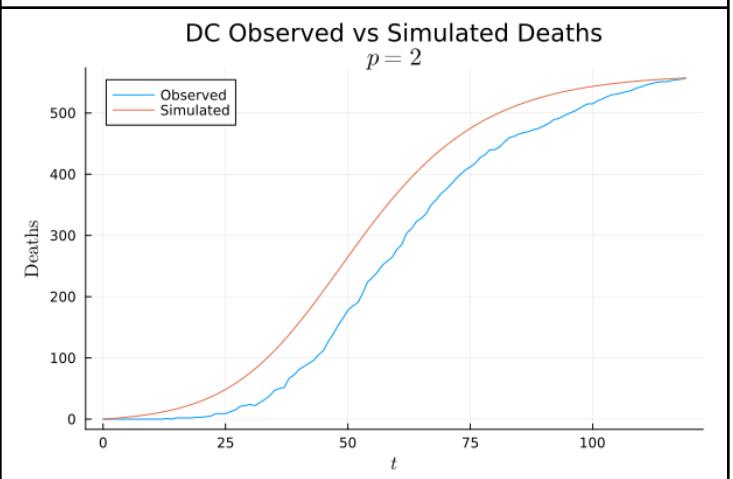
Here are the results for the SIR model fit to the first 120 days of the data on the DC Validation dataset ($T_{max} = 119$):

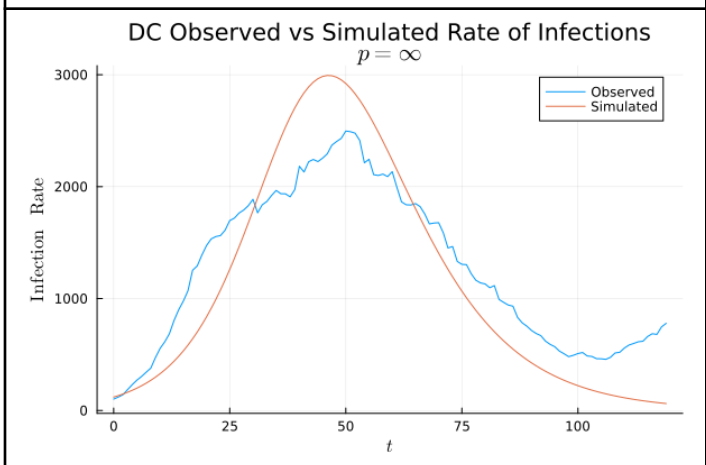
1 https://www.cdc.gov/mmwr/volumes/69/wr/mm6946a2.htm#:~:text=DOHMH%2C%202020).%E2%80%A0-,During%20February%2029%E2%80%93June%201%2C%202020%2C,-a%20total%20of

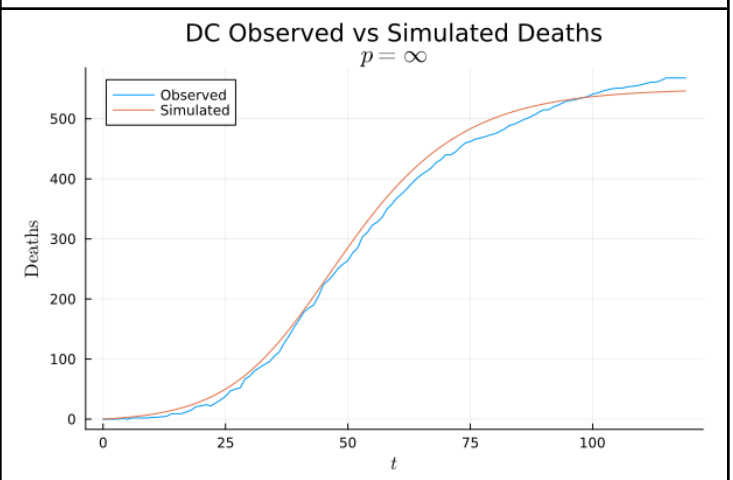












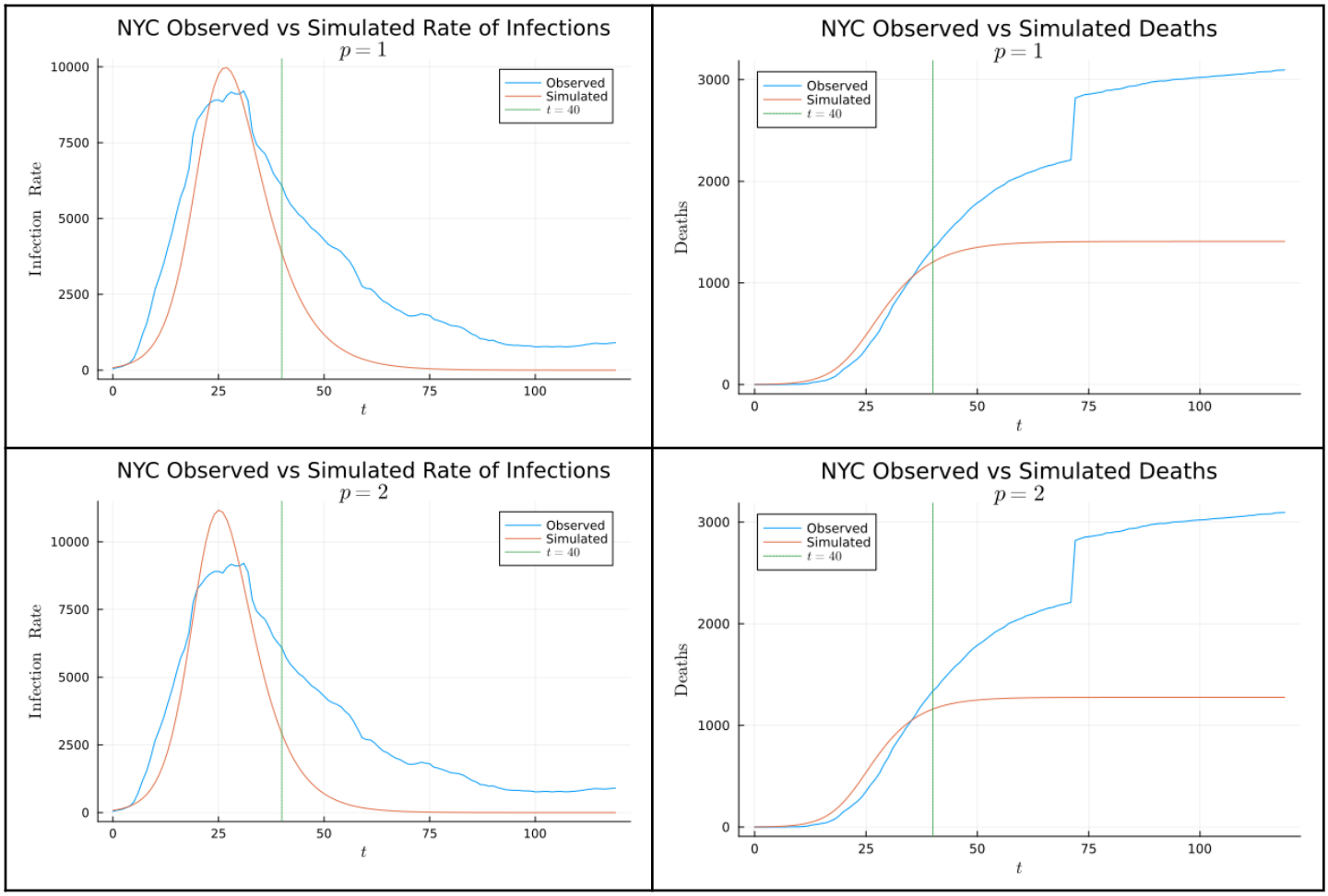
Parameters:

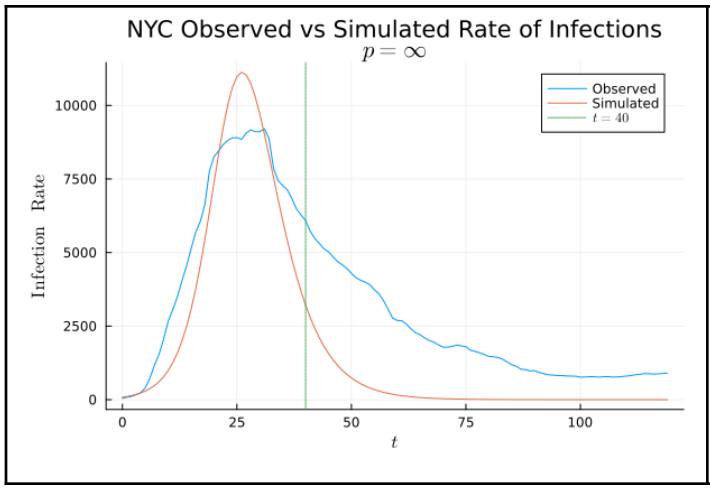
	p = 1	**P = 2**	**P = ∞**
α	0.1	0.12	0.13
β	0.19	0.216	0.234

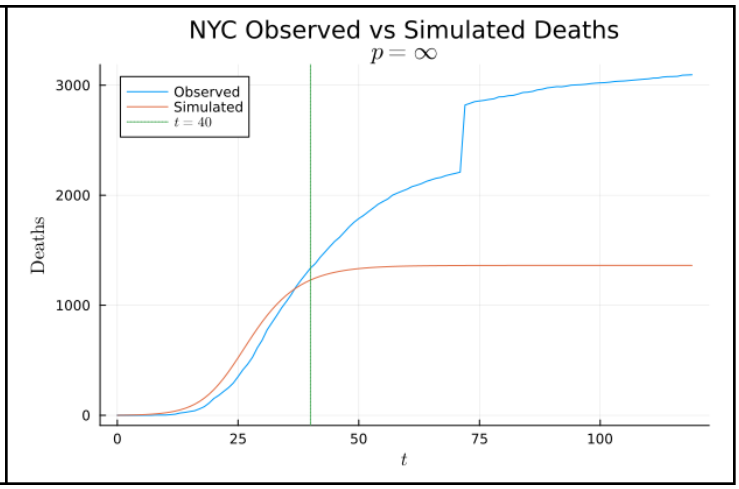
R_0	1.9	1.8	1.8
N / N_{max}	2.5%	3%	3.5%
γ	0.04	0.03	0.03
J	32002.7	1.39×10^7	738.9

It was interesting to see that deaths were modeled quite well and that none of the models was very close to modeling the real infections. It is also interesting to see the difference in gamma that we obtain in DC vs NYC. (DC much lower)

Here are the results for the SIR model fit to the first 40 days of the data on the Manhattan dataset ($T_{max} = 39$):



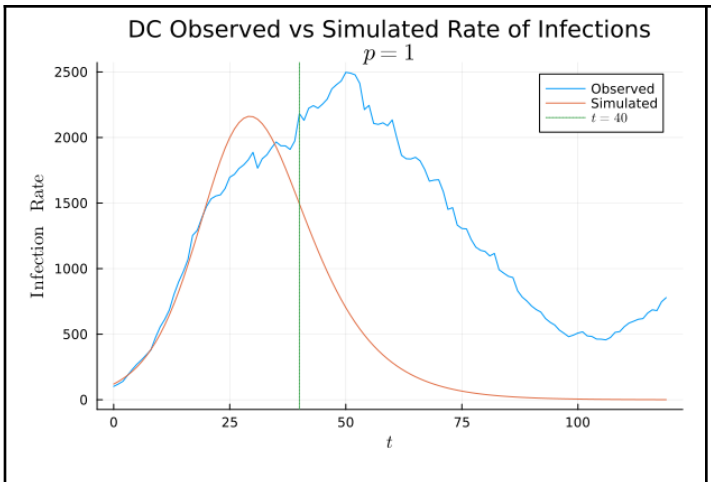


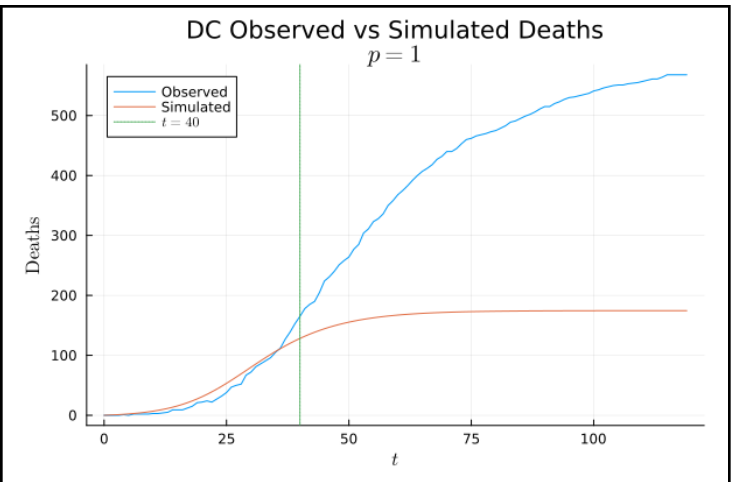


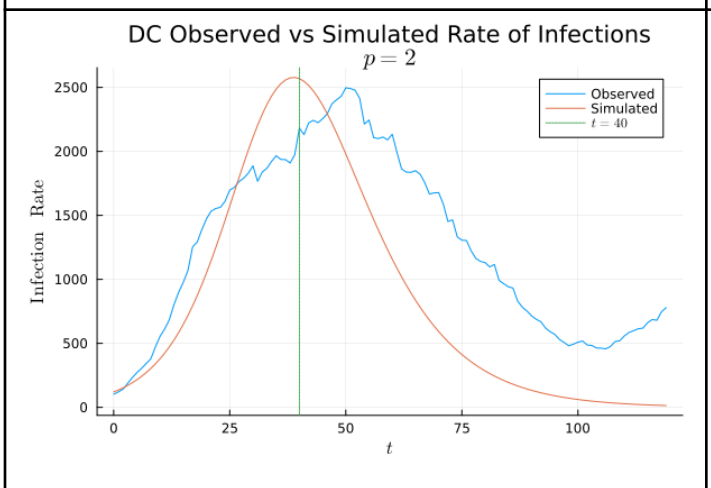
Parameters:

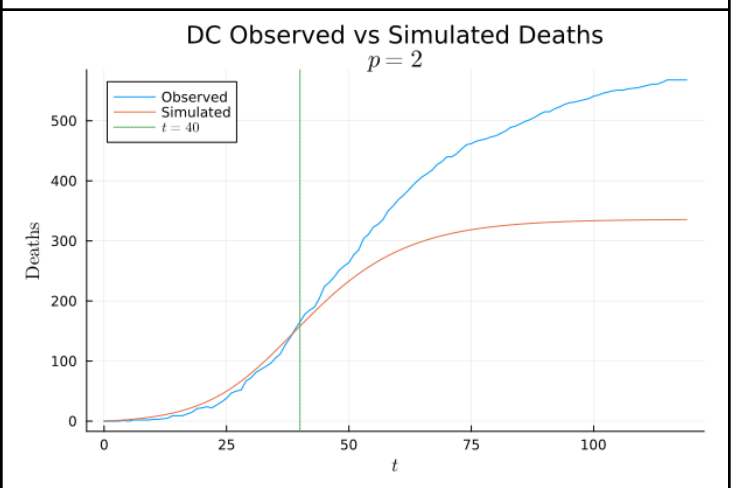
	p = 1	**P = 2**	**P = ∞**
α	0.19	0.24	0.28
β	0.44	0.504	0.532
R_0	2.3	2.1	1.9
N / N_{max}	3%	4%	5%
γ	0.05	0.02	0.02
J	49618.6	9.38×10^7	2846
t when I(t) < V_{min}	94 -> 4/24/2020	84 -> 4/14/2020	84 -> 4/14/2020

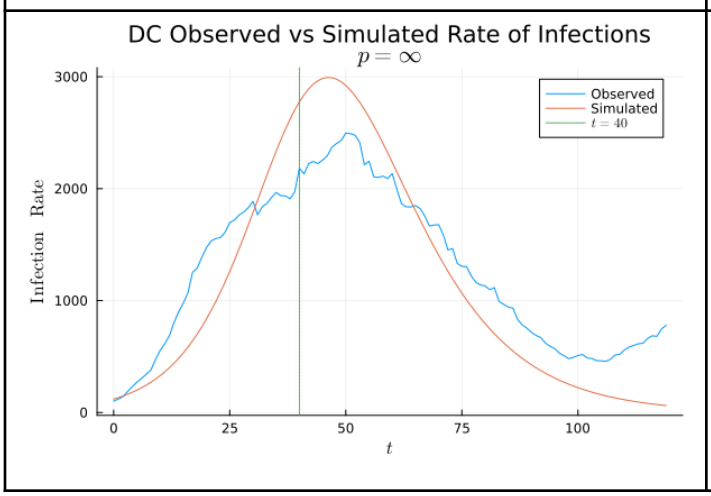
Here are the results for the SIR model fit to the first 40 days of the data on the DC validation dataset ($T_{max} = 39$):

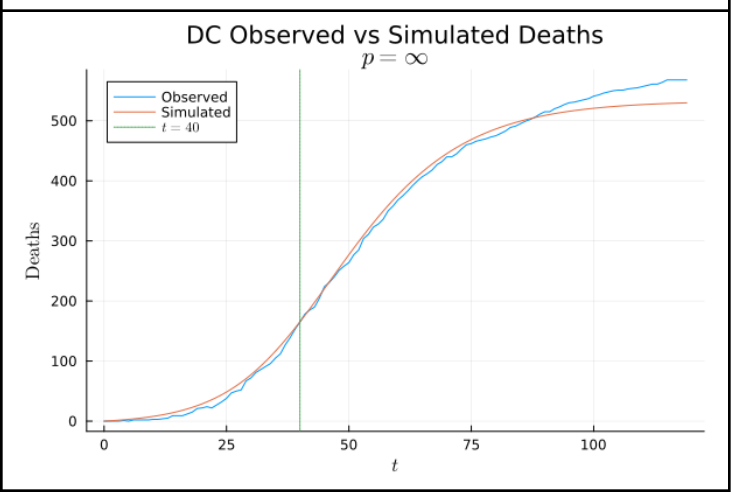

DC Observed vs Simulated Rate of Infections
$p = 1$
Observed
Simulated
$t = 40$
Infection Rate
t


DC Observed vs Simulated Deaths
$p = 1$
Observed
Simulated
$t = 40$
Deaths
t


DC Observed vs Simulated Rate of Infections
$p = 2$
Observed
Simulated
$t = 40$
Infection Rate
t


DC Observed vs Simulated Deaths
$p = 2$
Observed
Simulated
$t = 40$
Deaths
t


DC Observed vs Simulated Rate of Infections
$p = \infty$
Observed
Simulated
$t = 40$
Infection Rate
t


DC Observed vs Simulated Deaths
$p = \infty$
Observed
Simulated
$t = 40$
Deaths
t

Parameters:

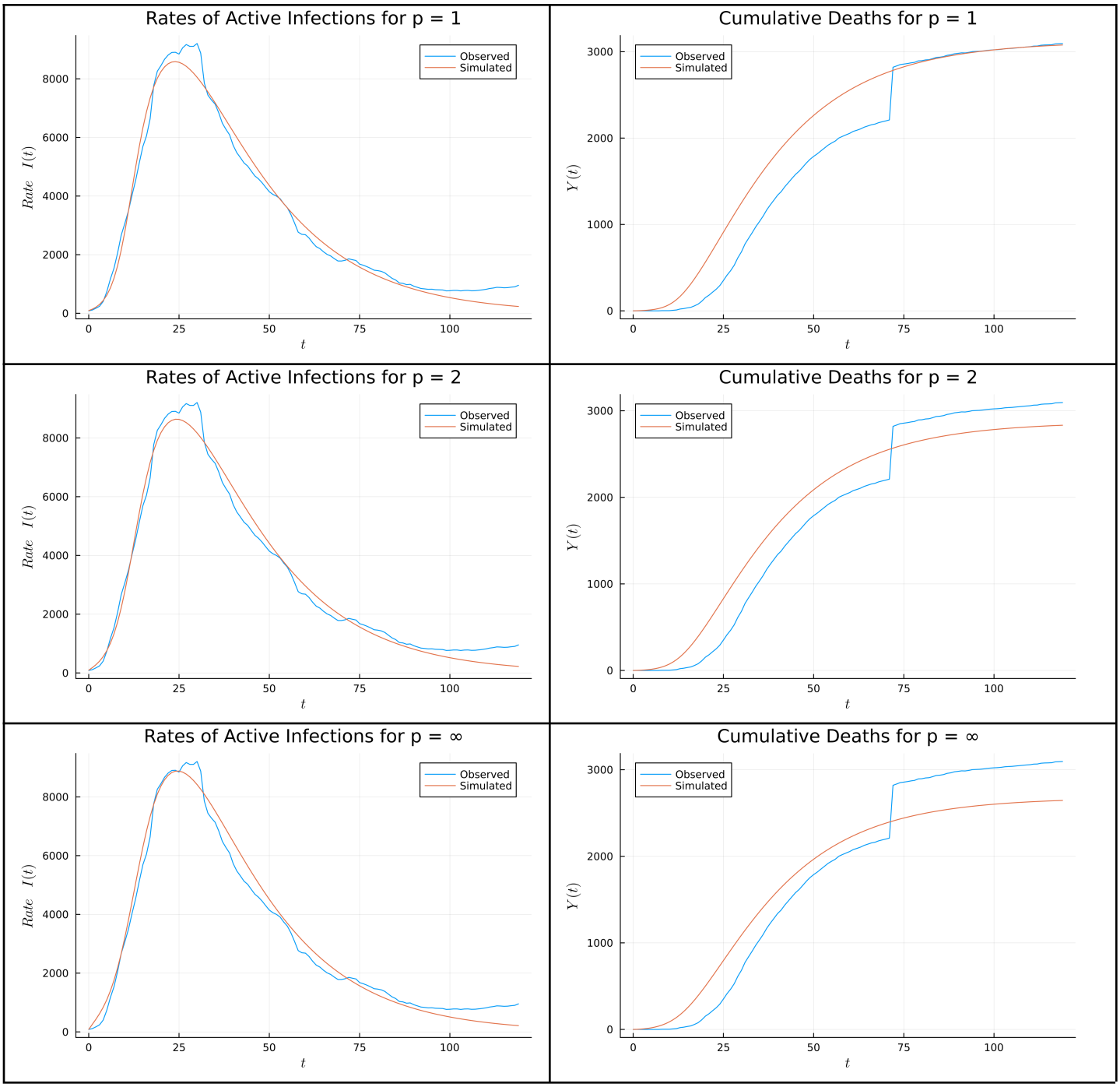
	p = 1	**P = 2**	**P = ∞**
α	0.19	0.15	0.13
β	0.342	0.27	0.234
R_0	1.8	1.8	1.8
N / N_{max}	2.5%	3%	3.5%
γ	0.013	0.021	0.029
J	6654.8	4.99×10^6	721
t when I(t) < V_{min}	102 -> 5/2/2020	Never	Never

It is really interesting to see how for p inf the deaths are modeled almost perfectly. There is some variation in the models for infections and all fail to accurately model infections.

For the SEIR models we increased the ranges for parameters searched and also introduced the $E_{sim}(0)$ and $I_{sim}(0)$ parameters. We also fixed δ at 0.1 since if left be it would optimize to be as high as possible. The new ranges for parameters are as follows:

- α ∈ [0.03, 0.4]
- R_0 ∈ [1.5, 50]
- δ = 0.1
- N / N_{max} $\in [0.005,\ 0.1]$
- E(0) / I(t_0) ∈ [1, 30]
- I(0) / I(t_0) ∈ [1, 30]

Here are the results for the SEIR model fit to the first 120 days of the data on the Manhattan dataset ($T_{max} = 119$):



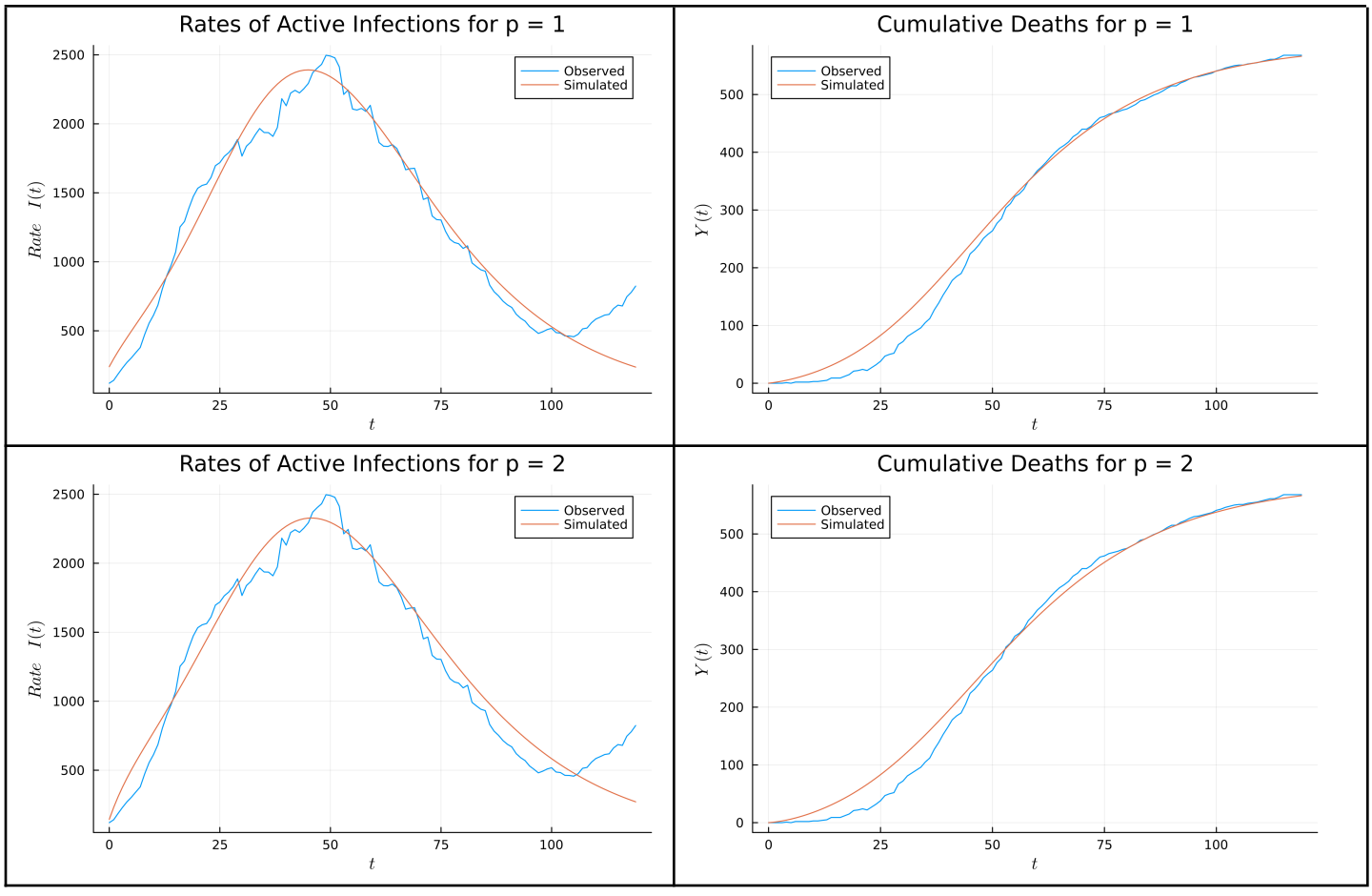
Parameters:

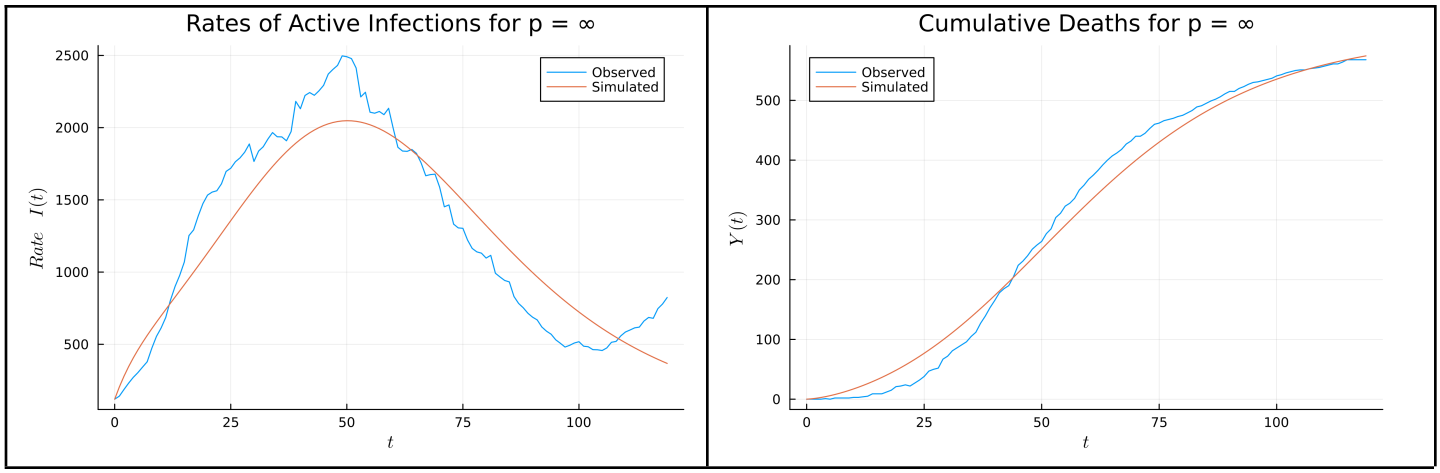
	p = 1	**P = 2**	**P = 3**
α	0.044	0.045	0.046
β	1.98	1.53	1.16
R_0	45	34	25

δ (fixed)	0.1	0.1	0.1
N / N_{max}	1%	1%	1%
γ	0.185	0.166	0.147
E_{sim}**(0) / I(0)**	5.6	10.8	25.2
I_{sim}**(0) / I(0)**	1.1	1.1	1
J	65,083	26,010,1112	1236

The model fit the infected curve very closely for all values of p, and close to the deaths curve except for being unable to fit the large jump in the death data. α was low at around 0.045 for all of them. R_0 was very higher and varied a between 25 and 45 for different p values. γ was pretty high, between 0.185 and 0.147. There were also a lot of people in the E category at t=0 with between 5 and 25 times the amount of I(0).

Here are the results for the SEIR model fit to the first 120 days of the data on the DC validation dataset ($T_{max} = 119$):



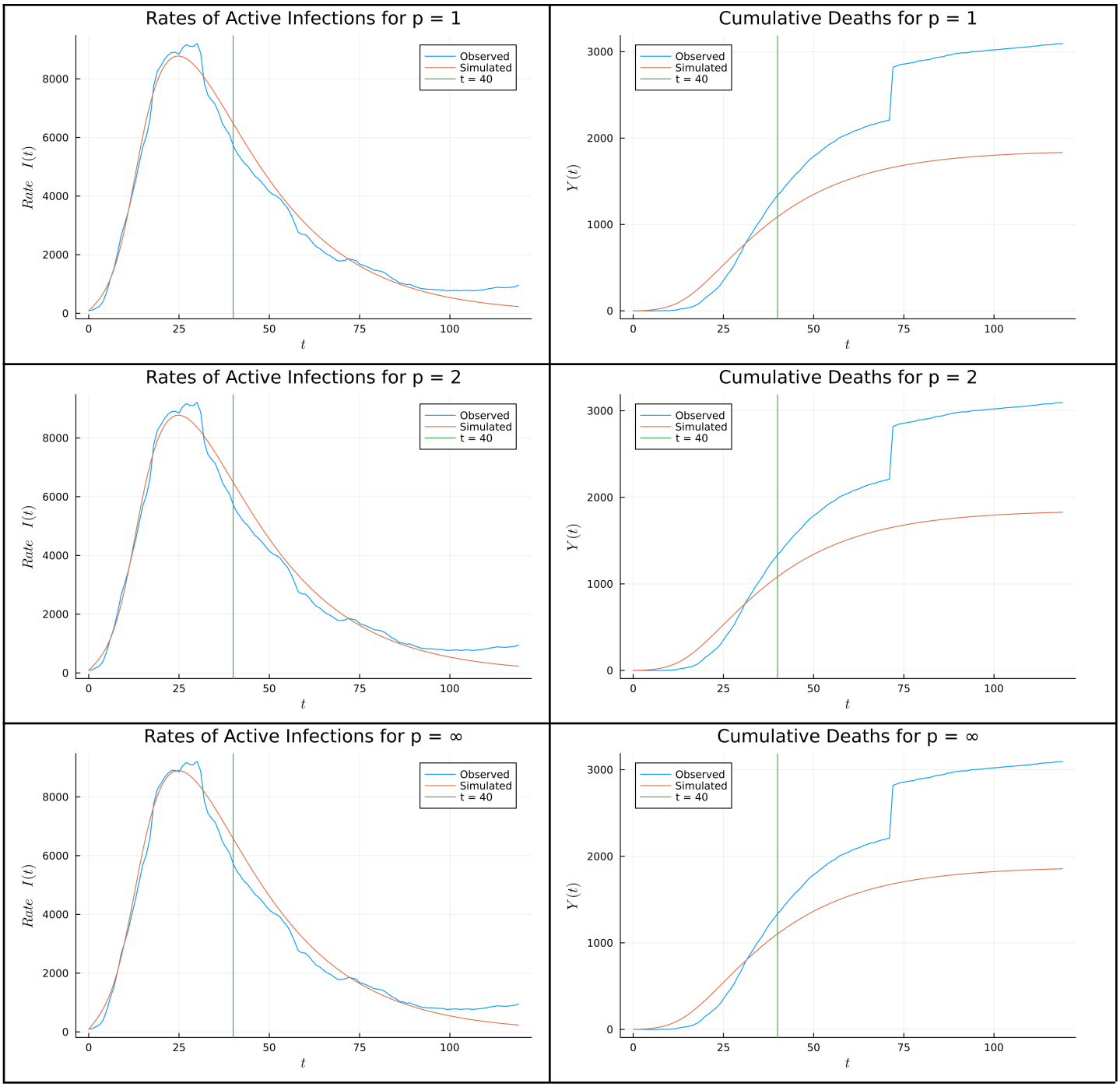


Parameters:

	p = 1	P = 2	P = 3
α	0.05	0.05	0.05
β	0.22	0.2	0.17
R_0	4.4	4.0	3.4
δ (fixed)	0.1	0.1	0.1
N / N_{max}	1%	1%	1%
γ	0.074	0.073	0.078
E_{sim}**(0) / I(0)**	6.0	8.3	8.0
I_{sim}**(0) / I(0)**	2.0	1.2	1.0
J	16,698	3,170,662	498

The model fit the curve very well in most cases except for not matching the increase near the end of the data. Only p = ∞ seemed to be underestimating the main peak by a bit. The deaths also matched very well except they are a bit higher than the observed deaths at the start. α was 0.05 for each p, R_0 was always around 4, and $E_{sim}(0)/I(0)$ was between 6 and 8.3. γ was around 0.075

Here are the results for the SEIR model fit to the first 40 days of the data on the Manhattan dataset ($T_{max} = 39$):



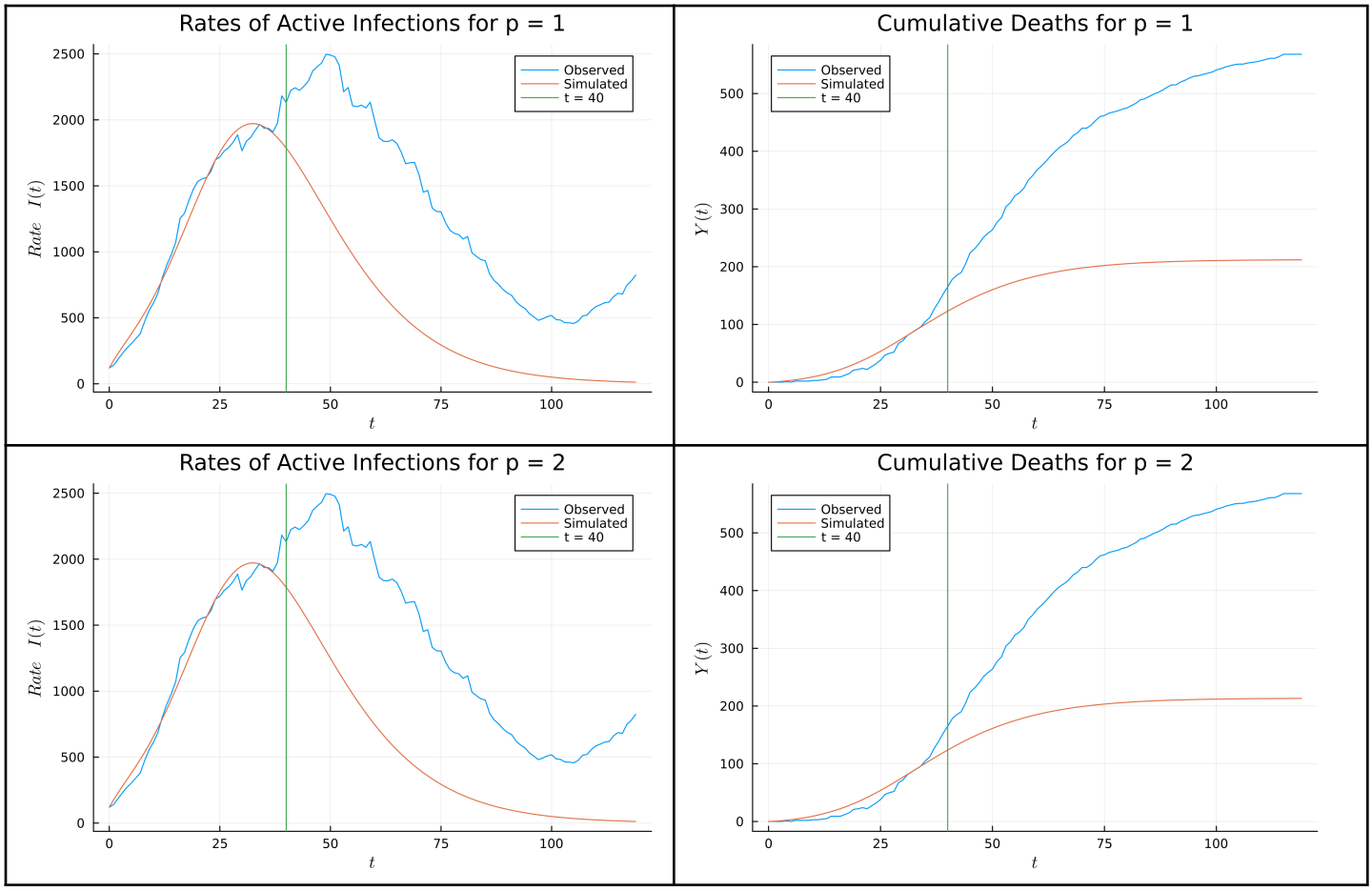
Parameters:

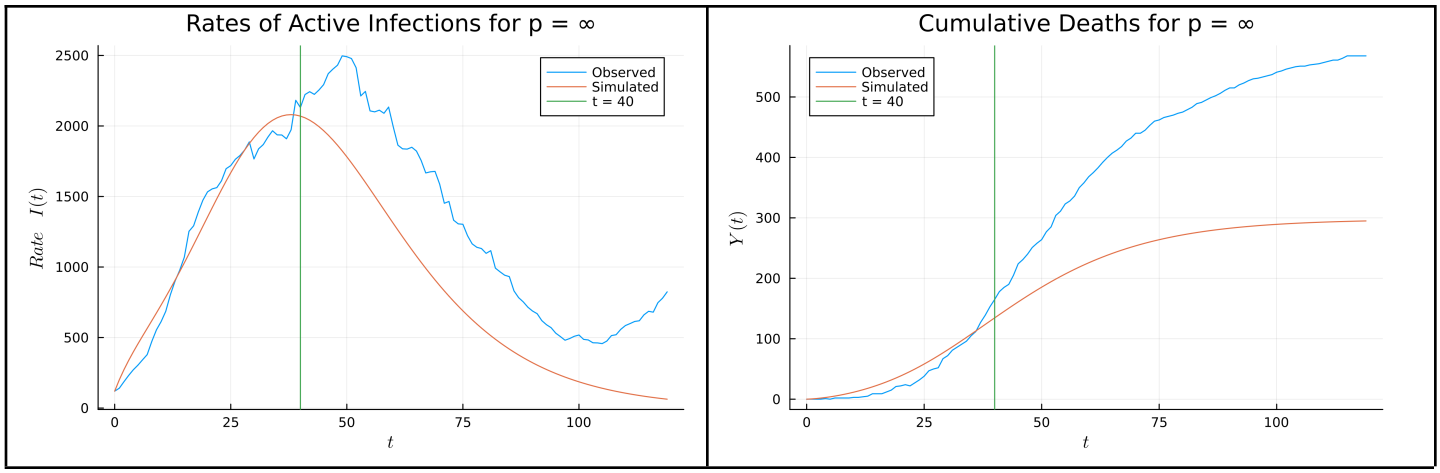
	p = 1	**P = 2**	**P = 3**
α	0.045	0.045	0.045
β	1.28	1.27	1.17
R_0	28.5	28.2	26.0

δ (fixed)	0.1	0.1	0.1
N / N_{max}	1%	1%	1%
γ	0.105	0.104	0.104
E_{sim}**(0) / I(0)**	16.1	16.1	20.1
I_{sim}**(0) / I(0)**	1.0	1.0	1.0
J	16,398	6,122,289	933
t when I(t) < V_{min}	206	206	206

The curve still fit the infected data very well. However, the deaths were underestimated by a lot. The parameters were very similar to the $T_{max} = 119$ case except for γ which was much lower at around 0.104 since the deaths were underestimated by a lot.

Here are the results for the SEIR model fit to the first 40 days of the data on the DC validation dataset ($T_{max} = 39$):





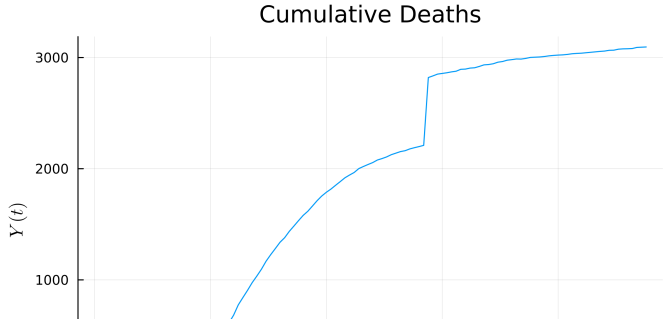
Parameters:

	p = 1	P = 2	P = 3
α	0.09	0.09	0.07
β	0.45	0.45	0.29
R_0	5.0	5.0	4.2
δ (fixed)	0.1	0.1	0.1
N / N_{max}	1%	1%	1%
γ	0.027	0.027	0.037
E_{sim}**(0) / I(0)**	6.0	6.0	7.9
I_{sim}**(0) / I(0)**	1.0	1.0	1.0
J	3,136	368,520	210
t when I(t) < V_{min}	132	132	163

The of infections peaks at the small dip around t=40 so the number of infections is underestimated. The number of deaths is also greatly underestimated. α is about twice as large as the t=119 case, around 0.09. The rest of the parameters are similar except for γ which is now much larger, around 0.075.

Discussion

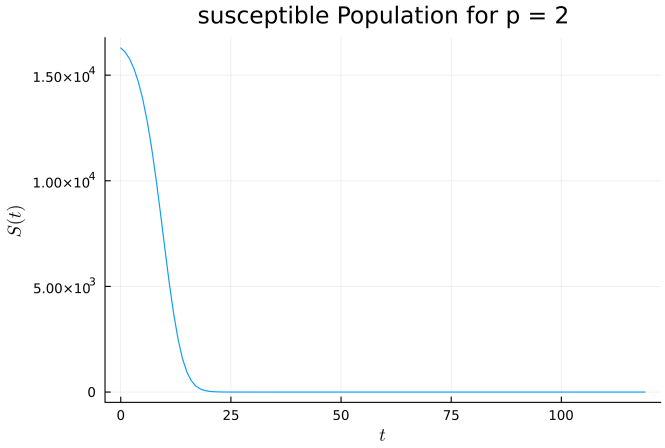
Our first goal for this problem was to fit the model to the first 120 days of the data in order to use the parameters of the model to estimate some properties of the virus. There were



some challenges in accomplishing this goal. One problem is that there is a large jump in the deaths data.

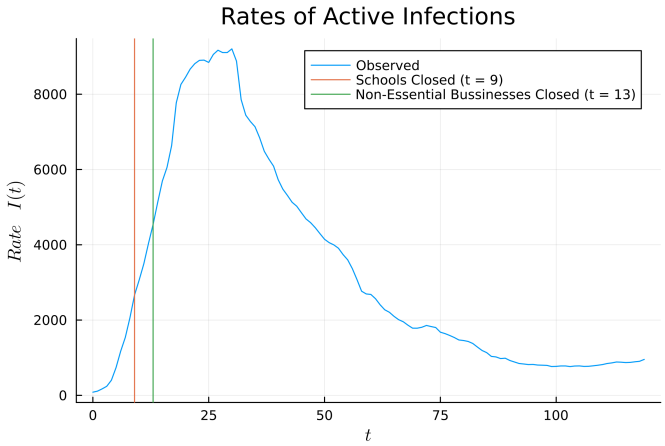
This was an increase of 610 deaths between May 17, 2020 and May 18, 2020. Such a large jump with no lead up before or after is unlikely to be just a normal fluctuation in the data and is most likely caused by a change in how they reported deaths. We were unable to find a definite reason for this jump in Manhattan. Similar jumps in data have been caused by switching from only counting people with positive PCR tests to also counting people with positive antigen tests which are cheaper and faster but less accurate (Pokin, 2021). Another possibility is that they changed criteria for which deaths are considered to be caused by covid or some new set of data was added to the total. In any case this could affect the accuracy of the death rate γ, but for the p = 1 case the data for SIR and SEIR seem to be fitting to the top part of the death curve so that gamma may match what it would have been if the whole data was recording using the criteria of the later part.

Another problem is that the SEIR model was fitting to data in a weird way. The β was well over 1 so people were very quickly moved out of the susceptible category.



So the peak in infected was caused by the population of susceptible people running out. This made the values of β and R_0 very high so they are unlikely to accurately describe properties of the virus when interpreted.

One explanation for why this weird match happens may be that the real β was not constant throughout. In the middle of the peak New York City instituted some major lockdown measures including closing all schools and closing all non-essential businesses (Governor Cuomo Signs executive order closing schools statewide for two weeks) (Governor Cuomo Issues guidance on essential services under the 'New York State on pause' executive order).



This likely caused a large change in β which is dependent on human behavior. Since these models only can have a constant β it makes sense that they aren't normally able to match the shape of the data.

This means that the interpretation of the values likely won't give any insight into the real properties of the virus. For the SEIR model in Manhattan these interpretations are an average infectious period (1/α) of 21.7 to 22.7 days and an average number of close contacts per infected individual (R_0) of 25 to 45. The only one that may be accurate is the death rate since the model did still match the number of infections well and fit to the death data well also. The range for this was 14.7% to 18.5%.

For the DC validation data, the SEIR model fared much better fitting the curve well and in a normal way. This means that the interpreted values may make sense. The average infectious period (1/ α) was 20 days. The average number of close contacts per infected individual (R_0) was 3.4 to 4.4, and the death rate was 7.4% to 7.8%. All of these seem like reasonable values. Another thing to notice is that the model had 6 to 8 times as many people in the exposed category as there were measured to be infected at that time. This may suggest that there were lots of sick people that weren't being tested at that time.

The second goal was to see if models fitted to only the first 40 days of the data would have any predictive capabilities. For the SEIR model and the Manhattan dataset the prediction fit the infections data very closely, however the peak of this data occurred within the first 40 days so this is not a very difficult prediction task since most of the key information is contained within those first 4 days. The deaths were very underestimated since to match the deaths later on the model had to increase away form the measured values initially but since it was only fitting the first 40 days it just increased at a much slower rate. The predicted time where the number of infections drops below $V_{min} = 5$ was t = 206, this is pretty high since the curve has that long tail trailing off.

For the SEIR model fitting the DC validation data the prediction was much worse for the infected. The model had its peak at a small dip that occurred around t = 40 so it underestimated the amount of infected by a decent amount. The number of deaths was also drastically underestimated. . The predicted time where the number of infections drops below $V_{min} = 5$ was t = 132 or t = 163 in the p = ∞ case. These models seem to be too simplistic to have much predictive power for these datasets.

Works Cited

Governor Cuomo Issues guidance on essential services under the 'New York State on pause' executive order. Governor Kathy Hochul. (n.d.). Retrieved March 16, 2023, from https://www.governor.ny.gov/news/governor-cuomo-issues-guidance-essential-services-under-new-york-state-pause-executive-order

Governor Cuomo Signs executive order closing schools statewide for two weeks. Governor Kathy Hochul. (n.d.). Retrieved March 16, 2023, from https://www.governor.ny.gov/news/governor-cuomo-signs-executive-order-closing-schools-statewide-two-weeks

Pokin, S. (2021, August 19). *Answer man: What's behind the big jump in county covid-19 cases reported in March?* Leader. Retrieved March 16, 2023, from https://www.news-leader.com/story/news/local/2021/08/19/answer-man-why-graph-shows-big-jump-greene-county-covid-19-cases/8184223002/