The SEIR model of COVID-19 forecasting rates of infections in New York City

Jiajing Zha¹, Xiangdong Liu¹,²,∗ and Minggang Sui¹
¹Department of Economics, Jinan University, Guangzhou, Guangdong, China
²Department of Statistics, Jinan University, Guangzhou, P.R. China

Corresponding author’s e-mail: tliuxd@jnu.edu.cn

Abstract. We used SEIR mathematical models to simulate the spread of COVID-19 in New York City, selecting a period of time to estimate the total number of infections and to predict the date of the highest number of infections. We found that the cumulative number of people infected or infected and the proportion of those infected increased with the increase in infectivity; The peak of infectious diseases is brought forward with the decrease of infectiousness; The number of people, \( I_e + I_u \) (prevalence), and the estimated peak of the disease occurred on April 25. At \( R_0 \), the regeneration number is much higher than 1, which is proportional to the infection rate. Since the infectiousness of unreported cases is not known, we can predict the outbreak under three different scenarios and get reasonable results.

1. Introduction
This report presents the models and results used to study COVID-19 transmission in New York City on 17 February 2020 solstice on 31 July 2020. From the surveillance data, we know the number of reported cases. However, it is also important to know the number of unreported cases, as this will help to predict, plan and evaluate possible interventions in the future.

2. Objective
The goal of the model is to estimate the total number of COVID-19 infections in New York City, including unreported cases, and to estimate peak days, when the number of infections is highest and the number of new cases is highest.

3. NCP propagation model - SEIR model[1]

3.1. Assumptions
We developed a so-called compartment model in which individuals are divided into compartments based on predefined properties. Within each region, individuals are assumed to have the same characteristics and to have the same behaviour characteristics. Where S stands for susceptibility, E for exposure, I for infection and R for recovery. When a healthy person is infected, he or she does not immediately become infectious, but instead enters asymptomatic Stage E and stays in that compartment for an average of 5.1 days[2].

We divided the infections into two groups: reported cases and unreported cases. After the incubation period, the infected person is either tested in the health care sector and becomes a confirmed and reported case or remains undiagnosed, i.e., unreported case. If COVID-19 is confirmed...
and reported, an individual will be transferred from room E to \( I_r = I_{\text{reported}} \). If the individual is untested and unreported, he or she will be transferred from Room E to \( I_u \). The unreported probability is \( p_u \), and the reported probability is \( P_r = 1 - p_u \). Suppose contagiousness, that is, the contagiousness rate varies between the values \( \theta \) and \( \delta \theta \), where the midpoint between \( \theta \) and \( \delta \theta \) occurs at the time \( t_b \). Let's say the tipping point occurs on April 15, when people in New York City are advised to work from home. The rate of infectious change is determined by parameters, which increase or decrease depending on the combination of \( \varepsilon \) and \( \delta \). A contagious disease is described as follows\[3]\: \( b_t = b(t, t_b, \theta, \delta, \varepsilon) = \theta \left( \delta + \frac{1 - \delta}{1 + e^{-\varepsilon(t - t_b)}} \right) \). The special case which \( \varepsilon = 0 \) will lead to persistent infection rate.

We hypothesized that the infectivity of individuals who become reported cases follows the time dependence of infection rate, and the infectivity between unreported cases is the factor of infectivity between reported cases \( q_0 \in [0, 1] \). Assume that a person is contagious for an average of 5 days and moves to room R after 5 days\[4\]. Suppose New York City is a lockdown, where no one enters or leaves the crowd. The number of individuals in each compartment is denoted by \( S,E,I_r,I_u,R \). This indicates a population size of\[5\]

\[ N = S + E + I_r + I_u + R \]  

(1)

3.2. Model setup

Transmission dynamics is described by the following equation system:

\[ \frac{dS}{dt} = -S \frac{b_t I_r}{N} - S q_0 \frac{b_t I_u}{N} \]  

(2)

\[ \frac{dE}{dt} = S \frac{b_t I_r}{N} + S q_0 \frac{b_t I_u}{N} - \rho E \]  

(3)

\[ \frac{dI_0}{dt} = p_0 \rho E - \gamma I_u \]  

(4)

\[ \frac{dI_r}{dt} = (1 - p_0) \rho E - \gamma I_r \]  

(5)

\[ \frac{dR}{dt} = \gamma (I_u + I_r) \]  

(6)

From this differential equation system, the number of initial individuals in different compartments at the desired starting time \( t_0 \), we can calculate the Numbers \( S,E,I_r,I_u,R \) at any time \( t \). It is worth noting that the number of reported cases per day in the model is \( (1 - p_0) \rho E(t) \), which has been calibrated to accommodate the actual number of reported cases per day in New York City.

3.3. Reproductive number

The basic reproductive number \( R_0 \)\[6\] is defined as the expected number of people infected by an initial infected person in a fully susceptible population. In our model, \( R_0 \) varies over time because it is assumed that the infectivity rate depends on time, which is given by multiplying the infectivity rate by the expected infectivity time of the initially infected individual; Infectivity rate refers to the number of people infected per unit of time by an infectious individual. In our model, infected individuals are either reported cases with \( b_t \) infection or unreported cases with low \( q_u b_t \) infection. In terms of unreported ratios, we get

\[ R_0(t) = (1 - p_u)b_t / \gamma + p_0 q_u b_t / \gamma \]  

(7)

To obtain the effective reproductive number \( R_e(t) \), multiply \( R_0(t) \) times the proportion of susceptible individuals at time \( t \), i.e.,

\[ R_e(t) = R_0(t) S(t) / N \]  

(8)

4. The fitting instructions

4.1. Fit the observation data
We matched the parameters of the SEIR model with the number of domestic cases reported daily in New York City as of 31 July 2020. More specifically, we fitted the model parameters so that the estimated incidence \((1 - p_0)\rho E(t)\) of reported cases per day was similar to the observed incidence of reported cases per day. The length of time the infected person was infected and the duration of the incubation period were determined, we estimated all parameters \(\theta, \delta\) and \(\epsilon\) for the different infectiousness. Infected cases abroad were excluded because they were not infected in that population.

We use the date on which symptoms are reported as the date on which the case is contagious. Based on the data collected, we assume that there are no unreported cases before 2 March. Therefore, we assume that at time \(t_0 = 2\) March, 1 case of infection was reported and no unreported case. Note that we set the number of reported cases to be \(t_0\) and the number of unreported cases to be \(\frac{p_u}{p_r} = \frac{p_u}{1-p_u}\).

When fitting the parameters into the data, we minimized the residual square error between the number of new cases per day estimated by the model and the number of new cases observed per day.

4.2. New York Area Health Report[7,8]

The New York department of public health conducted a study that sampled covid-19 in the area during July 27, solstice, and August 3, 2020. The results of this study were used to calibrate our model and estimate the number of unreported cases.

The results show an average prevalence of 1.6 per cent during 27 July solstice 3 August 2020.

5. Methods

5.1. Fit the model

We fitted the mathematical SEIR model[9] to different assumptions of \(q_u\) parameter values, and \(q_u\) was a factor to describe the relationship between the infectivity of unreported cases and reported cases. Three different conditions were calibrated so that the estimated prevalence of infected individuals (i.e., \(I_r + I_u\)) averaged 1.6% between 27 July solstice and 3 August. According to our calculations, if the prevalence was 1.6% on a given date, the probability of unreported cases was \(p_u = 0.97\). This means that 97 percent of those infected are unreported cases.

5.2. Three levels of infectivity

We then investigated three different levels of infectivity in unreported cases relative to reported cases: \(q_u = 0.5, 0.8, 1.1\). These values were chosen at random, but based on the basic assumption that infectivity is related to the degree of symptoms. Thus, it is reasonable to assume that milder cases are less contagious than heavier ones. Based on the above differential equation system, we conclude that these scenarios will primarily affect the relative infectivity of unreported cases to reported cases. We estimate the model parameters and their standard errors under three conditions.

In order to obtain the guided confidence interval (CI) of the parameters, such as the estimated curve of the number of cases reported in the new day, we drew 1000 combinations of model parameters, each parameter was the parameter estimation from the normal distribution, and was used as its standard error mean and standard deviation and curve by calculating the differential equation. This method is used to estimate confidence intervals for all results in the table.

| Parameters                                                                 | Value                      |
|----------------------------------------------------------------------------|----------------------------|
| Length of incubation period \(\frac{1}{\rho}\)                           | 5.1 days                   |
| Length of infectiousness \(\frac{1}{\gamma}\)                           | 5 days                     |
| Start date \(t_0\)                                                       | 2 March                    |
| Turning point between \(\theta\) and \(\delta \theta\) occurs at time.   | 15 April                   |
| Probability that a case remains unreported                                | 0.97 (calibrated)          |
| Reduction factor of the infectiousness of an unreported case compared to a case, \(q_u\) | Scenarios with \(\{0.5, 0.8, 1.0\}\) |
| Infectivity rate parameters \(\theta\)                                   | Estimated by the model     |
| Infectivity rate at time \(b_i\)                                        | Estimated by the model     |
6. Results of forecasting

Note that as a general principle, for large tables font sizes can be reduced to make the table fit on a page or fit to the width of the text. Based on cases reported in New York City as of July 31, 2020[10], we studied three scenarios. The results of all three scenarios are consistent with the assumption that the COVID-19 prevalence was about 1.6% on 27 July solstice on 3 August. What all three scenarios have in common is that unreported cases must account for 97% of the total number of cases, and unreported cases are not the same as reported case ($q_u$) infectivity.

6.1. Peak days and number of infections[11,12]

According to the scenario investigated, the peak epidemic day, the date with the highest number of simultaneous infections, occurred on 8 solstice 11 April, when about 70,500 people were infected. The highest number of new cases, morbidity or inflow $I (I_r + I_u)$, occurred on 2 or 3 April, when about 14,500 people were infected.

Table 2 shows the cumulative number and proportion of people infected or infected ($E + I + R$) under three different conditions. On August 1, 1.5 percent of the population in New York City was infected or had been infected. Under the three scenarios, the cumulative number of persons increases with the increase of Infectiousness.

Table 2. Cumulative number of infected persons and share ($E + I + R$) for the two dates. Three common scenarios are where a case has a 0.97 probability of becoming an unreported case ($p_u$), but assume that the infectivity of an unreported case is different from that of a reported case ($q_u$). All estimates are expressed in terms of 95% CI.

| Scenario | 2020-08-01 | 2020-08-11 |
|----------|------------|------------|
|          | Number (95%CI) | Proportion (95%CI) | Number (95%CI) | Proportion (95%CI) |
| 100%     | [150139, 18913726] | [0.008, 0.984] | [161864, 18933828] | [0.008, 0.985] |
| 80%      | [133826, 662754] | [0.007, 0.015] | [143685, 731311] | [0.007, 0.038] |
| 50%      | [195792, 18720359] | [0.01, 0.974] | [207854, 18775176] | [0.011, 0.977] |

Under different conditions, the peak days of coinfection (i.e., all individuals with $I_r + I_u$) were shown in Table 3. The peak day of infectious disease has been advanced by the decrease of infectiousness; the number of persons, $I_r + I_u$ (prevalence rate), and the estimated peak value of morbidity occurred on April 25.

Table 3. Peak days and number of infectious diseases estimated by unreported cases relative to reported cases ($q_u$) infectivity, $I_r + I_u$ (prevalence), estimated incidence peak days, number of new cases per day, i.e., daily inflow $I_r + I_u$.

| Scenario | Prevalence | Incidence |
|----------|------------|-----------|
|          | Peak day (95%CI) | Prevalence on peak-day (95%CI) | Peak day (95%CI) | Prevalence on peak-day (95%CI) |
| 100%     | 2020-08-16 | 12486     | 2020/4/15 | 2970     |
|          | [2020-06-06,2020-08-31] | [6270, 3888408] | ----- | -----     |
| 80%      | 2020-08-15 | 12656     | 2020/4/15 | 2978     |
|          | [2020-05-09,2020-08-31] | [5758, 32044] | ----- | -----     |
| 50%      | 2020-08-13 | 12908     | 2020/4/15 | 2970     |
|          | [2020-06-03,2020-08-31] | [8340, 3608838] | ----- | -----     |
In Figures 1, 2, and 3, the left shows the estimated progression of newly reported cases, daily incidence, or $I_r$ influx, while the right shows the total number of infected individuals, both reported and unreported, with a confidence interval of 95%.

(a) The infectivity is 100%

(b) The infectivity is 80%

(c) The infectivity is 50%

Figure 1. Fitted SEIR model under different scenarios

We assume that infectivity between unreported cases is 80% of reported cases.

- Left: Estimated daily incidence of newly reported cases, $I_r$ influx cases, confidence interval 95% (dotted line) and observed data (circle).
- Right: Prevalence of infectious cases, both reported and unreported, with a confidence interval of 95%.
7. Forecast
The results of the three scenarios are predicted as follows. The data of specific variables in the model are shown in the table below:

Figure 2. Infectivity between unreported cases is 80% of reported cases.

Figure 3. Fitted reported cases including prediction.
Figure 4. Estimated reproductive number and estimated infectivity under different scenarios.

Based on the cases reported during 2 March solstice and 31 July, we calculated the regeneration value, $R(t)$ at a certain point. In these three scenarios, the regeneration number at the moment of $R(0)$ is much
higher than 1, and is proportional to the infectivity. \( R_0 \) accepts a value slightly above or slightly below 1 on July 31. The number of copies of the 95% confidence interval is shown in Table 4.

**Table 4.** Cases in our analysis report on the last day of July 31, 2020 under the 95% confidence interval estimate effective reproductive number (\( R_e \)) and relative to the unreported cases reported cases of infectious (\( q_u \)).

| \( q_u \) | \( R_0(2020-03-02) \) (95%CI) | \( R_0(2020-07-31) \) (95%CI) | \( R_e(2020-07-31) \) (95%CI) |
|---|---|---|---|
| **Scenario 1:** 100% | 4.101 | 1.017 | 0.655 |
| | [0.988, 4.344] | [0.986, 4.302] | [0.068, 1.025] |
| **Scenario 2:** 80% | 4.011 | 1.017 | 1 |
| | [0.98, 1.055] | [0.969, 1.03] |
| **Scenario 3:** 50% | 3.823 | 1.013 | 0.579 |
| | [0.961, 3.905] | [0.961, 3.902] | [0.103, 1.042] |

In Table 5 we present the estimated parameters and the residual squared error (RSS) in three cases. It shows three scenarios that provide the best fit for the data. In scenario 1, for example, the peak epidemic in coinfect populations occurred on 16 August (95% confidence interval 6 June solstice August 31). The prevalence of coinfection was 12,486, with a 95% confidence interval [6270, 3888408]. Although the distribution of infectivity was different between reported and unreported cases, there was little difference in the number and duration of infection under different scenarios.

**Table 5.** Estimates the prevalence of unreported cases as a percentage of the New York City population as of 3 August 27 solstice, the sum of squares of residuals (RSS) and the standard error (se) and confidence interval (95% CI) of estimated parameters, based on the infectivity of unreported cases relative to reported cases (\( q_u \)).

| \( q_u \) | Prevalence | \( RSS \) (95%CI) | \( \hat{\delta} \) (95%CI) | \( \hat{\varepsilon} \) (95%CI) | \( \hat{\theta} \) (95%CI) |
|---|---|---|---|---|---|
| **Scenario 1:** 100% | 1.6% | 0.00131 | 4.033 | [3.807, 4.258] | 9.852 | [62.653, 82.536] | 0.203 | [0.197, 0.21] |
| | | | | | | 325301.198 | 0.115(se) | 36.993(se) | 0.004(se) |
| **Scenario 2:** 80% | 1.6% | 0.00130 | 3.944 | [3.686, 4.202] | 7.04 | [6.813, 7.266] | 0.252 | [0.243, 0.262] |
| | | | | | | 324088.7 | 0.132(se) | 0.115(se) | 0.005(se) |
| **Scenario 3:** 50% | 1.6% | 0.00129 | 0.265 | [0.25, 0.28] | -30.64 | [-441.753, 380.472] | 1.485 | [1.449, 1.521] |
| | | | | | | 321313.117 | 0.008(se) | 209.755(se) | 0.018(se) |

8. Limitations
COVID-19 is mainly transmitted through droplet infection, suggesting that the social contact structure of the population has an important influence on infection dynamics. The compartment model does not take into account the differences in interpersonal contact that occur in a society where few people have a lot of contact and most people have even less. This simplification in the model, the uniform contact structure, usually results in a slightly faster rate of epidemic growth than if the model included heterogeneity. The model therefore carries a risk of the speed of the outbreak in the New York area. This is not included in the specified confidence interval because the confidence interval does not report such uncertainty.

We assume that the infection rate (infectivity) varies over time, but the parameters controlling the incubation period (\( \rho \)) and the infectivity (\( \gamma \) length are constant. That means they control the random variable of the exponential distribution. This is a (common) simplification of the partitioning model
for ease of calculation, but it is worth emphasizing. The time of exponential distribution is not always the best description of what we observe. For example, the change in the incubation period may be lower than the change brought about by the exponential distribution time with an average of 5.1.

Furthermore, we assume that the relationship between reported and unreported cases is constant over time. This is a limitation, as the sampling routine was changed on 15 April 2020, when people were banned from gathering and wearing masks when going out, and medical resources were strengthened. We believe that this change has not significantly affected the reporting of domestic infections, but rather the reporting of previously imported cases.

**Acknowledgment**

This author would like to thank a referee and an Associate Editor for their comments and suggestions.

**References**

[1] Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epidemic processes in complex networks. Rev Mod Phys. 2015;87:925-79. doi:10.1103/RevModPhys.87.925

[2] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med. 2020; 382:1199-207. Medline:31995857 doi:10.1056/NEJMoa2001316.

[3] Public Health Agency of Sweden, Year 2020. Article number: 20059.

[4] Cao Z, Zhang Q, Lu X, Pfeiffer D, Song H, Pei T, et al. Incorporating human movement data to improve epidemiological estimates for 2019-nCoV. Med Rxiv. 2020;2020.02.07.20021071.

[5] Haken H. Laser theory. In: Light and matter Ic/Licht und materie Ic. Springer; 1970. p. 1–304.

[6] Zhou T, Liu Q, Yang Z, Liao J, Yang K, Bai W, et al. Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV. J Evid Based Med. 2020; 13:3-7. Medline:32048815 doi:10.1111/jebm.12376.

[7] https://www.nyc.gov/site/doh/covid-19-data-testing.page access:2020.9.28.

[8] Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science (2020) doi:10.1126/science.abb3221.

[9] Wan, Jing Chen, Changming Lu, Lanlan Dong, Zhicheng Wu, Lianglu Zhang. When will the battle against novel coronavirus end in Wuhan: A SEIR modeling analysis. www.jogh.org doi: 10.7189/jogh.10.011002.

[10] Johns Hopkins Coronavirus Resource Center (CRC) https://coronavirus.jhu.edu/ access:2020.9.29

[11] Read JM, Bridgen JR, Cummings DA et al. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. medRxiv. 2020;2020.01.23.20018549.

[12] Rodríguez-Morales AJ, Cardona-Ospina JA, Gutiérrez-Ocampo E, Villamizar-Peña R, Holquín-Rivera Y, Escalera-Antezana JP, et al. Clinical, laboratory and imaging features of COVID-19: a systematic review and meta-analysis. Travel Med Infect Dis 2020:101623.