COVID-19 Virus Prediction of US

001051249 Yu Ren
001089477 Jing Bian
001056959 Jinnuo Che

Introduction

Since the end of 2019, a new type of coronavirus pneumonia has erupted in China, and the epidemic has swept the country. It is another national spread of viral pneumonia after SARS. Now, outbreaks are happening all over the world, we plan to analyze the trend of the epidemic objectively in America so as to prevent and avoid the virus effectively. We will use SIR model to model the number of changes of confirmed and recovered people in China. This model will contain two important parameters, namely, β (rate of infection per susceptible and per infective individual) and γ (rate of recovery) parameters. But because β should change with the time, so we find another way to model the problem that is the Logistic model. In this model, we could set a value r to represent resistance which could change with time. So the β could be more accurate. Based on social comparisons, we can estimate American β parameter so that we can predict what the situation will be in USA.

Methodology

1. Problems Analysis

The prediction of epidemic situation can be regarded as infectious disease science. In fact, most of the uploaded disease models are dynamic models (ordinary differential equations). With the dynamic model, there must be the identification of model parameters. The identification of model parameters can often be modeled as an optimization problem, and the solution of the optimization problem is just related to operation research

a. What do we know about nondrug interventions? For example, data sets on policy actions taken by regions (school closures, cancellation of large gatherings, self-isolation, etc.). Data that may help to estimate possible compliance with policy measures (e.g., small business concentration) and fear levels (which may be extracted from Google Trends or social media data).
b. What do you know about communication, incubation and environmental stability? For example, temperature, humidity and air pollution data by region.

c. What do we know about covid-19 risk factors? For example, the percentage of people who smoke in a region or country.
d. What publications about health care? For example, data on the number of doctors, nurses, hospitals by region or country.

2. Solutions

2.1 Introduce of SIR Model

SIR model is a common mathematical model to describe the spread of infectious diseases. Its basic assumption is to divide the population into the following three categories:

a. Susceptible: refers to the person who does not have the disease, but lacks the immune ability and is easy to be infected after contact with the patient.
b. Infectious: refers to the person infected with the disease, who can spread it to the susceptible population.
c. removed: the person removed from the system. A person who recovers (has immunity) or dies from illness. These people are no longer involved in the process of infection and being infected.

When the susceptible population contacts with the infected person, the infection rate is beta. The infection rate reflects the transmission intensity of the disease. The higher the infection rate is, the more likely the susceptible people will be infected after contact with the infected people.

The infected population recovered or died at a fixed average rate. The recovery rate gamma depends on the average duration of the infection.

For simplicity, the initial S, I, R of three groups of people are used to represent the number of three groups of people, and the N is used to represent the total number of labor. Then the rules of the dynamic change of the number of three groups of people with time can be expressed by the following ordinary differential equations:

$$\frac{dS}{dt} = -\beta\frac{SI}{N} \quad (1)$$

$$\frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I \quad (2)$$

$$\frac{dR}{dt} = \gamma I \quad (3)$$

2.1.1 SIR Model with nonlinear programming

SIR model uses a dynamic model (three ordinary differential equations) to model the process of three groups of people changing with time. Infection rate and recovery rate are used to quantify the behaviors of disease infection and disease cure. It is very important to obtain the accurate parameters of the dynamic model in order to build a relatively accurate model. So the main problem of using SIR model to model the spread of new pneumonia in Wuhan is to determine the following parameters

a. Infection rate Bata and recovery rate gamma
b. Initial value of susceptible population, infectious population, and removed population

Since the first case was confirmed on December 8, the initial time was chosen to be December 8. The initial value of the infected population is 1, the initial value of the susceptible population is *N-1*, the initial value of the removed population is 0, and *N* is the total population of Wuhan. Gamma is the novel coronavirus pneumonia, with a recovery rate of about 14 days, so *gamma = 1/14*. So here we focus on how to identify the accurate infection rate beta

In order to facilitate parameter identification, we simplify the above SIR model. We think that there is *S = N* in the early stage of disease transmission (the number of patients in the early stage of transmission is small, so we can approximately think that all people are susceptible population). We can get this condition from equation (2)

$$\frac{dI}{dt} = (\beta - \gamma)I \tag{4}$$

It is easy to know that the general solution of the differential equation is as follows:

$$I(t) = Ce^{(\beta-\gamma)t} \tag{5}$$

*C = 1*can be obtained by *I(t = 0) = 1*, and can be obtained by bringing it into formula (5)
Thus, the following parameter identification problems can be constructed:

$$I(t) = e^{(\beta-\gamma)t} \tag{6}$$

Decision variable: infection rate *beta*

Objective function:

$$\min \sum_{t \in T} \left(e^{(\beta-\gamma)t} - \hat{I}\,(t) \right)^2 \tag{7}$$

Where I is the actual number of patients (from the actual data), T is the time set, in days.
By solving the above optimization problem, we can get the infection rate *beta* of Wuhan New pneumonia. It is easy to know that the optimization problem is a non-linear and nonconvex optimization problem

2.1.2 SIR Model with pymc3

We will model observations with a log-normal distribution, a continuous probability density function (pdf) whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then *Y = ln(X)* has a normal distribution.
Taking the logarithm of a dataset is a big-time prettified of data and is often a first target model. So we will assume that our Chinese *I* (infected) and *R* (recovered) data are log-normally distributed. We will add dead people to the recovered population since that is how we built our infectious model above.
We will assume that there is noise/error in the Chinese data, so we will tack on a standard deviation. Since that quantity is never negative, it is often modeled as a Half Cauchy distribution. It has an interesting history.

2.2 Introduce of Logistic Model

The logistic function was proposed by Pierre Francois Verhulst, a Belgian mathematician and biologist when he studied the population growth model. It is an improvement on the Malthus population model (1798). The difference between the two models can be seen directly in the figure below. The two models are described in detail below.

Pierre improved the Malthusian population model by setting the population growth rate to r(1 − P / k), where k can be understood as the maximum population allowed by the environment. At this time, when the population P is closer to K, the growth rate is lower, that is, the population growth rate decreases linearly with the increase of population.

$$\frac{\mathrm{d}P}{\mathrm{d}t} = r(1 - \frac{P}{K})P$$

By solving the differential equation, we can get the function of the population changing with time,

$$P(t) = \frac{K}{1 + (\frac{K}{P_0} - 1)e^{-rt}}$$

2.2.1 Logistic Model with curve fit

Considering the heterogeneity between Hubei Province and other provinces, we will use the logistic model to fit the cumulative number of confirmed cases. First of all, it is necessary to obtain the cumulative number of confirmed cases. The data source is the wind data interface provided by windquant. For more data download methods, see n methods for obtaining the historical data of covid-19 epidemic situation.
Then the curve fit function of scipy. An optimized library is used to fit the logistic curve. The undetermined parameters include K, p_andr. according to the minimum MSE criterion, the grid parameter adjustment method is used to find the optimal parameters: for the maximum capacity K, step 1 traverses (10000, 80000) interval, and for the growth rate R, step 0.01 traverses (0, 1) interval.

Dataset

1. Specifications of Dataset

The data is stored in the version library in CSV format while providing JSON and EXCEL format data

2. Introduction of the dataset

“COVID19-China.csv” is a dataset containing 70354 pieces of data about the virus situation changes in different cities and provinces in China. Each data contains province information, city information and the numbers of confirmed, suspected, cured and dead people in this city and this province. The data comes from the national, provincial, and Wuhan Municipal Health and Health Commission epidemic situation notice

字段	说明
date	时间（天）
country	国家
countryCode	国家代码
province	省
provinceCode	省代码
city	市
cityCode	市代码
confirmed	确诊人数
suspected	疑似人数
cured	治愈人数
dead	死亡人数

the description of columns

3. Dataset link

https://www.kaggle.com/wang749/china2019ncov

Results and Analysis

1. Nonlinear programming

1.1 Analysis of Nonlinear programming

The differential equations for the SIR model of infection are:

$$\frac{dS}{dt} = -\beta SI \quad S(0) = S_0$$

$$\frac{dI}{dt} = \beta SI - \gamma I \quad I(0) = I_0$$

$$\frac{dR}{dt} = \gamma I \quad R(0) = R_0$$

The quantity β/γ is called R-Nought (*R0*). Its interpretation is that if we were to drop a *single* infected person into a population of susceptible individuals, we would expect R0 new infections. If *R0*>1, then an epidemic will take place. If *R0*≤1 then there will be no epidemic.

When *R0*<1, each person who contracts the disease will infect fewer than one person before dying or recovering, so the outbreak will fizzle (*dI/dt*<0). When *R0*>1, each person who gets the disease will infect more than one person, so the epidemic will spread (*dI/dt*>0).

R0 is the most important quantity in epidemiology.

$$\boldsymbol{\beta = k * infection\_probability}$$

- Preprocess data
- Use Minimize to compute the Infection Probability
- Change the number of individuals that are contacted by each infected individual per day.

1.2 Evaluation of Results

- Result analysis: The number of contacts has a great influence on the trend of the epidemic, so quarantine is very important in this situation. In real life, the epidemic has not spread as in the model, mainly because national policies and quarantine have effectively reduced R0 in the later period, so the number of infected people in the later period of the epidemic will show a decreasing trend.

- Deficiencies: In the SIR model, β is obtained from the infection probability obtained after data analysis and the assumed k value. This β is not accurate enough. Can we process the known data and obtain a more accurate R0 to analyze the severity of the epidemic?

2. pymc3

2.1 Analysis of pymc3

If we know R(t) and I(t) then we can determine S(t): S(t)=1−I(t)−R(t), so we can work only with the two unknowns: R(t) and I(t). We prefer to work with these because that is what the China Covid19 dataset gives us!

$$S(t) + I(t) + R(t) = 1 \quad , \forall t$$

So we can write as:

$$\frac{dI}{dt} = \beta(1 - I - R)I - \gamma I \quad I(0) = I_0$$
$$\frac{dR}{dt} = \gamma I \quad R(0) = R_0$$

Simplify

$$\frac{dI}{dt} = \beta(1 - I - R - \gamma/\beta)I \quad I(0) = I_0$$
$$\frac{dR}{dt} = \gamma I \quad R(0) = R_0$$

- Preprocess data
- Use Differential Equation to create SIR model
- Assume that our Chinese I (infected) and R (recovered) data are log-normally distributed and use the Bayesian model to compute the parameters.
- Use the computed parameters to present the SIR modal

2.2 Evaluation of Results

- Result analysis: The parameter results we obtained after data processing are relatively accurate, and the drawing of these parameters conforms to the real situation.
- Deficiencies: The SIR model predicts the epidemic according to the given β and γ, and the population parameters of this epidemic model need to be set by ourselves. It only considers the spread of the epidemic but does not consider the resistance factors in reality. Can we find a model with consideration of resistance?

3. Curve - Fit

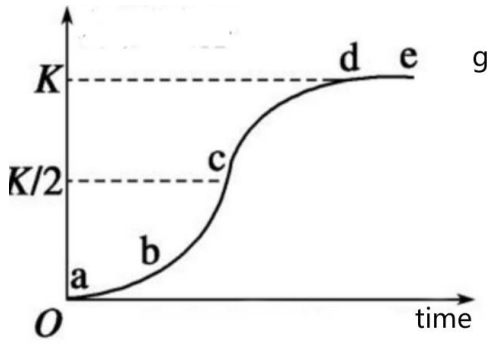
3.1 Analysis of Curve - Fit

1. Initial: density growth is slow due to the small number of individuals in the population
2. Acceleration: density growth accelerates as the number of individuals increases
3. Transition: density increases fastest when the number of individuals reaches half of the saturation density (K / 2)
4. Deceleration: growth slows down when the number of individuals exceeds half of the density (K / 2)
5. Saturation: number of individuals in the population reaches K value and is saturated

$$P(t) = \frac{KP_0e^{rt}}{K + P_0\left(e^{rt} - 1\right)} = \frac{K}{1 + \left(\frac{K - P_0}{P_0}\right)e^{-rt}},$$

where

$$\lim_{t\to\infty} P(t) = K.$$



1. Preprocess data
2. Use Curve Fit to compute capacity, intitial_value, and increase_rate
3. Use Logistic Model to show the trend of the confirmed population
4. Show the specific number of confirmed people every day in the future.

3.2 Evaluation of Results

- Result analysis: According to the prediction curve and daily detailed results, compared with the new data in recent days, our prediction is more accurate. Especially in recent days, the data is more practical. We can now see that with resistance set to linear growth, the number of confirmed cases will reach 709k by the end of April (and we hope so). I hope that government policies and people's behavior can make the resistance grow in a straight line as expected.

- Deficiencies: We set the initial value and growth rule of resistance by ourselves, so there are some subjective factors in predicting the trend of epidemic situation. In these two aspects, we may also use machine learning to make reasoning and judgment.

Conclusion

1. Current Conclusion

A. We know the trend of epidemic prediction displayed by Sir model under different R0, which let us know that human contact plays a decisive role in the trend of epidemic transmission.
B. We analyze the real R0 according to the known data, combined with the SIR model we have shown, we can know that the covid19 virus is a huge challenge for all human beings.
C. Based on the analysis of existing data and artificial judgment, we predict the rising curve of the number of confirmed cases in the United States under the assumption that the resistance increases linearly. We also hope that, as we predicted, the U.S. epidemic will gradually stabilize at the end of April, and we hope that the U.S. can get through the difficulties

2. Future

A. For SIR model, we may be able to dynamically analyze the change of R0 and predict its trend, so that Sir model can be more realistic.
B. In addition, for the initial number of people in SIR model, we can directly use the corresponding population instead, combining with the idea in 1, fitting a more real curve.
C. Because of the time, we only use one curve for the logistic model to speculate the resistance. Maybe we can use different curves and compare the results to find a more accurate one

Reference

1. https://www.kaggle.com/wang749/china2019ncov
2. https://blog.csdn.net/source_code13/article/details/104164299
3. https://blog.csdn.net/weixin_44533530/article/details/104732733
4. https://blog.csdn.net/z_ccsdn/article/details/104134358