COVID-19 : Statistical exploration

Ayoub Abraich[1]

[1]*Master 2 Data Science, Paris Saclay university . Email: ayoub.abraich@etud.univ-evry.fr*

Abstract

In this article we present a naive model for the prediction of the number of COVID-19 virus infections, with illustrations of real data on the evolution of COVID-19 in France. For more details, and for more complex models, see (Siettos and Russo 2013).

Keywords: COVID19,Machine Learning,prediction

1 Introduction

The concealed and apparently unpredictable nature of infectious diseases has been a source of fear and superstition since the first ages of human civilization. The worldwide panic following the emergence of SARS and COVID-19 in the world are examples that our feeling of dread increases with our ignorance of the disease . One of the primary aims of epidemic modeling is helping to understand the spread of diseases in host populations, both in time and space. Indeed, the processes of systematically clarifying inherent model assumptions, interpreting its variables, and estimating parameters are invaluable in uncovering precisely the mechanisms giving rise to the observed patterns.The very first epidemiological model was formulated by Daniel Bernoulli in 1760 with the aim of evaluating the impact of variolation on human life expectancy. However, there was a hiatus in epidemiological modeling until the beginning of the twentieth century 1 with the pioneering work of Hamer and Ross on measles and malaria, respectively.The past century has witnessed the rapid emergence and development of a substantial theory of epidemics. In 1927, Kermack and McKendrick derived the celebrated threshold theorem, which is one of the key results in epidemiology. It predicts – depending on the transmission potential of the infection – the critical fraction of susceptibles in the population that must be exceeded if an epidemic is to occur. This was followed by the classic work of Bartlett , who examined models and data to expose the factors that determine disease persistence in large populations.(Siettos and Russo 2013)

One of the best known models is SIR model . It's an epidemiological model that computes the theoretical number of people infected with a contagious iliness in a closed population over time. The name of this class of models derives from the fact that they involve coupled equations relating the number of susceptible people $S(t)$, number of people infected $\mathcal{I}(t)$, and number of people who have recovered $R(t)$. One of the simplest SIR models is the Kermack-McKendrick model (See).

2 Statistical modeling

2.1 Naive model

Let $N(t)$ be the number of cases infected at time t (day). We have the following recurrent (simplistic) formula which describes the dynamics of the spread of the epidemic:

$$N_{n+1} = N_n + E \cdot p \cdot N_n$$

with :

- p : the probability that the patient contaminates those who are in contact with him, it is strongly linked for example to the barrier gestures adopted by people.
- E : the average number of people who are in contact with the infected patient.

E	p	N_{10}	N_{50}
0.1	0.5	3	22
2	0.2	5.70e+01	4.05e+07
5	0.1	1.15e+02	1.28e+09
10	0.6	2.05e+03	2.25e+15
100	0.8	2.43e+19	5.31e+95

Table 1. Comparison of numbers of infections for different E and p.

At first, we consider E and p constant with respect to time (which is generally false!), but this will allow us to have an idea on the dynamics of COVID19. We obtain :

$$N_{n+1} = N_n + E \cdot p \cdot N_n \Leftrightarrow N_n = (1 + E \cdot p)^n N_0 = (1 + E \cdot p)^{n-n_0} N_{n_0}$$

We present in this table 1 a comparison of numbers of infections for different E and p to get an idea on the order of magnitude evolution.

We see that E and p determine the speed of propagation, which justifies the measures taken by the governments (ban on assembly in order to decrease E) and barrier gestures (in order to decrease p).

2.2 Improved model

In our simplistic model, we forgot (apart from the dependence over time of E and p) the fact that p decreases in time and tends to 0 : in fact, more the virus is spread , the probability of passing it on to someone who is already infected is almost zero. So, we have to take this into account, setting:

$$p(t) := 1 - \frac{N(t)}{n_{pop}}$$

with n_{pop} the number of population in the cluster studied or more generally in the whole country.

This implies a new simple differential equation (always assuming that E is independent of time t, which is generally wrong but practical):

$$N_{n+1} - N_n = E \cdot p \cdot N_n \rightarrow N'(t) = E \cdot \left(1 - \frac{N(t)}{n_{pop}}\right) N(t)$$

with $N(t_0) = N_0$

We can write this EDS as:

$$N'(t) = g(t, N(t))$$

with $g(t, y) := E \cdot (1 - \frac{y}{n_{pop}})\, y$ and g satisfies the conditions of the Cauchy-Lipschitz theorem,hence the existence and uniqueness of the solution.

The solution is exhibited in the form of logistics, ie:

$$N(t) = \frac{a}{\alpha + b \cdot \exp(c \cdot t)} = \frac{\tilde{a}}{1 + \exp(-\tilde{c} \cdot (t - \tilde{b}))}$$

We find by a simple calculation:

$$N(t) = \frac{n_{pop}}{\alpha + \gamma_0 \cdot \exp(E \cdot (t_0 - t))}$$

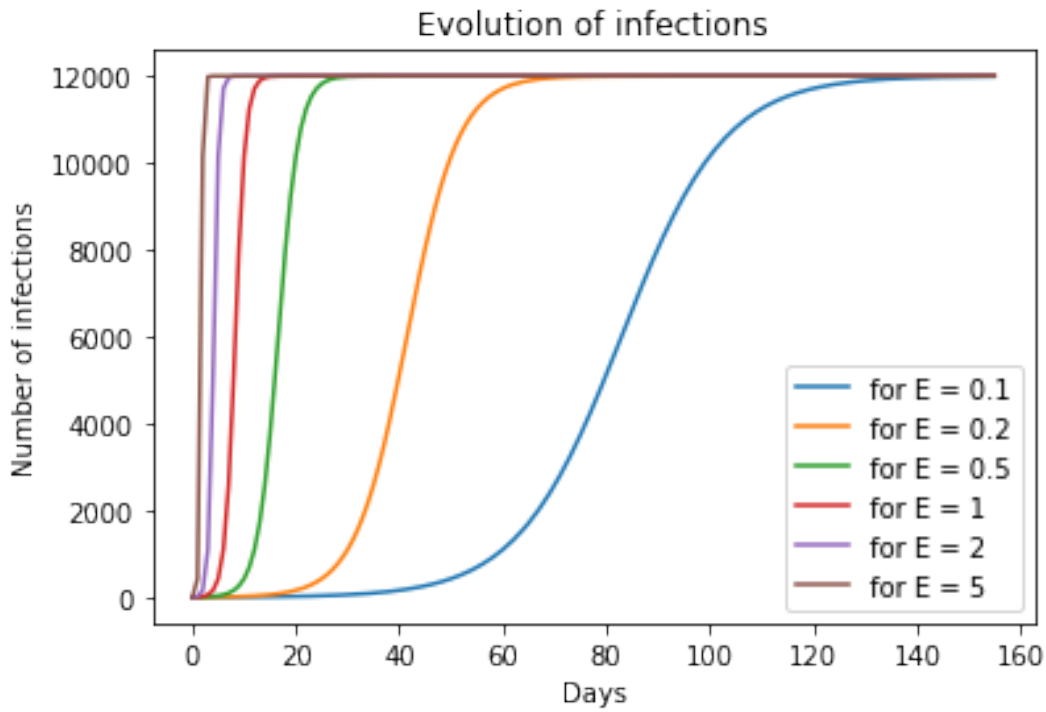

Figure 1. Evolution of infections for different E, $n_{pop} = 6 \cdot 10^7$, $\alpha = 5000$ and $N_0 = 2$.

with $\gamma_0 := \frac{n_{pop}}{N(t_0)} - \alpha$. Likewise for p:

$$p(t) := 1 - \frac{N(t)}{n_{pop}} = 1 - \frac{1}{\alpha + \gamma_0 \cdot \exp(E \cdot (t_0 - t))}$$

We denote :

$$\psi_t(\theta) := \frac{\theta_1}{1 + \exp(-\theta_3 \cdot (t - \theta_2))}$$

3 Machine learning

Now, we have to calculate $\theta^* := (\theta_0^*, \theta_1^*, \theta_2^*)$ such as

$$\theta^* = argmin_\theta \sum_{i=1}^{n} \left| X_{t_i} - \psi_{t_i}(\theta) \right|^2$$

with Gradient Descent methods for example (See). Then we define the predictor of $N(t)$ by :

$$\hat{N}(t) := \psi_t(\theta^*)$$

You will find all the code (in Python) for this paper on my Github page. We train this model for France COVID-19 dataset, we get the total final number of cases expected in France (with 95.44% confidence interval) : Fig 2 We also tried to predict the number of infections and deaths in the short term, ie the highly exponential part of the curve using a simple GLM model.

References

Siettos, C., and L. Russo. 2013. "Mathematical modeling of infectious disease dynamics." *Virulence* 4 (April). doi:10.4161/viru.24041.

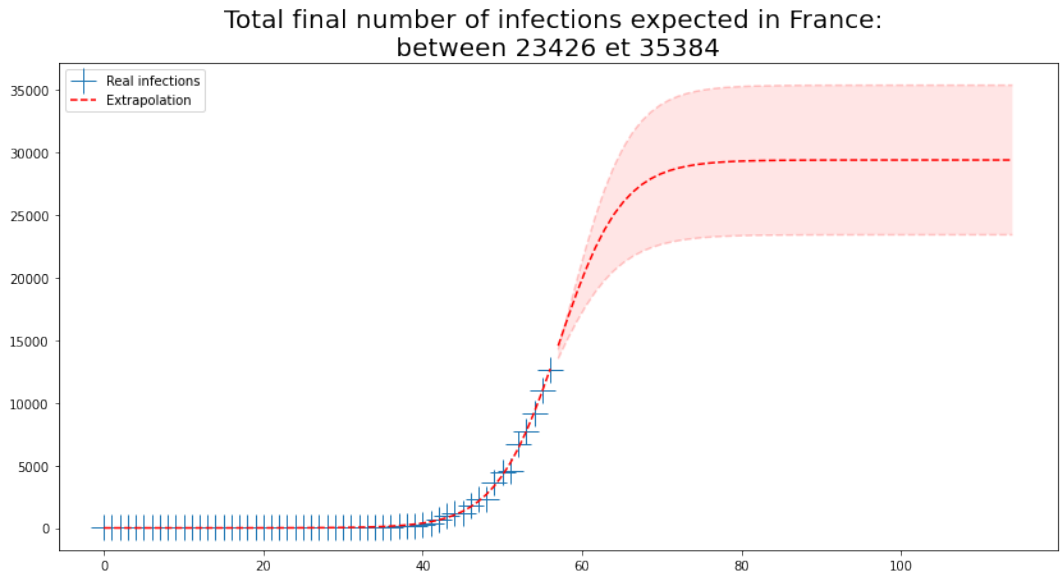

Figure 2. Total final number of cases expected in France