Estimation of the final size of the COVID-19 epidemic

Milan Batista
University of Ljubljana, Slovenia
milan.batista@fpp.uni-lj.si

(Feb 2020)

Abstract

In this short paper, the logistic growth model and classic susceptible-infected-recovered dynamic model are used to estimate the final size of the coronavirus epidemic.

1 Introduction

One of the common questions regarding an epidemic is its final size. To answer this question various models are used: analytical (Danby 1985, Brauer 2019a, b, Murray 2002), stochastic (Miller 2012), and phenomenological (Fisman D 2014, Pell et al. 2018).

In this note, we attempt to estimate the final epidemic size using the phenomenological logistic growth model (Pell et al. 2018, Chowell G 2014) and the classic susceptible-infected-recovered (SIR) model (Hethcote 2000). With both the models, we obtain a series of daily predictions. The final sizes are then predicted using iterated Shanks transformation (Shanks 1955, Bender and Orszag 1999). The data used for the calculations are taken from *worldmeters*[1].

Before proceeding, we note that the final size of the epidemic in its early stage was discussed by Wu et al. (Wu, Leung, and Leung 2020) using the susceptible-exposed-infected-resistant model, by Xiong and Yan (Xiong and Yan 2020) using the exposed-infected-resistant model, by Nesteruk (Nesteruk 2020) using the SIR model, and by Anastassopoulou et al. (Anastassopoulou et al. 2020) using the SIR/death model. These early predictions range from 65000 to a million cases. Roosa et al recently gave short-term forecasts of the epidemic (Roosa et al. 2020).

[1] https://www.worldometers.info/coronavirus/

2 Logistic growth model

The logistic growth model originates from population dynamics (Haberman 1998). The underlying assumption of the model is that the rate of change in the number of new cases per capita linearly decreases with the number of cases. Hence, if C is the number of cases, and t is the time, then the model is expressed as

$$\frac{1}{C}\frac{dC}{dt} = r\left(1 - \frac{C}{K}\right), \tag{1}$$

where r is infection rate, and K *is* the final epidemic size. If $C\left(0\right) = C_0$ is the initial number of cases, then the solution of (1) is

$$C = \frac{K}{1 + A\exp\left(-rt\right)}\ , \tag{2}$$

where $A = \dfrac{K - C_0}{C_0}$. The growth rate, $\dfrac{dC}{dt}$, reaches its maximum when $\dfrac{d^2C}{dt^2} = 0$. From this condition, we obtain that the growth rate peaks at time $t_{p.}$

$$t_p = \frac{\ln A}{r} \tag{3}$$

At this time, the number of cases and growth rate are

$$C_p = \frac{K}{2}, \qquad \left(\frac{dC}{dt}\right)_p = \frac{rK}{4}. \tag{4}$$

Now, if $C_1, C_2, \ldots, C_n$ are the number of cases at times $t_1, t_2, \ldots, t_n$, then the final size predictions of the epidemic based on these data are $K_1, K_2, \ldots, K_n$. By using Shanks transformation, the predicted final epidemic size is

$$K = \frac{K_{n+1}K_{n-1} - K_n^2}{K_{n+1} - 2K_n + K_{n-1}}. \tag{5}$$

For the practical calculation of the parameters K and r, we use the MATLAB functions *lsqcurvefit* and *fitnlm*.

3 SIR model

The model equations are

$$\frac{dS}{dt} = -\frac{\beta}{N} IS\,, \tag{6}$$

$$\frac{dI}{dt} = \frac{\beta}{N} IS - \gamma I\,, \tag{7}$$

$$\frac{dR}{dt} = \gamma I\,, \tag{8}$$

where t is time, $S(t)$ is the number of susceptible persons at time t, $I = I(t)$ is the number of infected persons at time t, $R(t)$ is the number of recovered persons in time t, β is the contact rate, and $1/\gamma$ is the average infectious period. From (1), (2), and (3) we obtain the total population size, N.

$$N = S + I + R = \text{const.} \tag{9}$$

The initial conditions are $S(0) = S_0$, $I(0) = I_0$, and $R(0) = R_0$.

Eliminating I from (1) and (3) yields

$$S = S_0 \exp\left[-\frac{\beta}{N\gamma}\left(R - R_0\right)\right]. \tag{10}$$

In the limit $t \to \infty$, the number of susceptible people left, S_∞, is

$$S_\infty = S_0 \exp\left[-\frac{\beta}{N\gamma}\left(R_\infty - R_0\right)\right], \tag{11}$$

where R_∞ is the final number of recovered persons. As the final number of infected people is zero, we have, using (4),

$$N = S_\infty + R_\infty\,. \tag{12}$$

From this and (6), the equation for R_∞ is

$$R_\infty = N - S_0 \exp\left[-\frac{\beta}{N\gamma}\left(R_\infty - R_0\right)\right]. \tag{13}$$

To use the model, we must estimate the model parameters β, γ, and the initial values S_0 and I_0 from the available data (we set $R_0 = 0$ and $I_0 = C_0$).

Now the available data is a time series of the total number of cases C, i.e.,

$$C = I + R\,. \tag{14}$$

We can estimate the parameters and initial values by minimizing the difference between the actual and predicted number of cases, i.e., by minimizing

$$\left\| C_t - \hat{C}_t\left(\beta, \gamma, S_0\right) \right\|^2 = \min, \tag{15}$$

where $C_t = \left(C_1, C_2, \ldots, C_n\right)$ are the number of cases at times $t_1, t_2, \ldots, t_n$ and $\hat{C}_t = \left(\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_n\right)$ are the corresponding estimates calculated by the model. For practical calculation, we use the MATLAB function `fminsearch`. For the integration of the model equation, we use the MATLAB function `ode45`.

With a series of predicted final number of recovered persons, $R_{\infty,1}, R_{\infty,2}, \ldots, R_{\infty,n}$, we can estimate the series limit by Shanks transformation.

$$R_\infty = \frac{R_{\infty,n+1} R_{\infty,n-1} - R_{\infty,n}^2}{R_{\infty,n+1} + R_{\infty,n-1} - 2R_{\infty,n}} \tag{16}$$

4 Results

The results of logistic regression and the SIR model simulation are given in Tables 1 and 2, respectively. The comparison of the predicted final sizes is shown in the graph in Figure 1. We see that both methods converge and with more data, the discrepancy between the predicted values becomes less than 5%. From Table 1, we see that the peak of the epidemic was probably on 9 Feb, 2020.

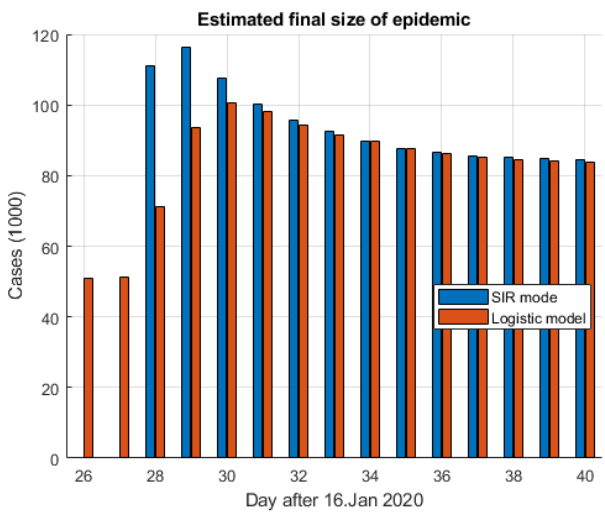

Figure 1. Evaluation of estimated final size of coronavirus epidemic (data until 20 Feb, 2020)

Table 1. Data and results of logistic regression (see Eqs. (2), (3), (4))

data			Regression			Peak		
date	day	C (cases)	K (cases)	r (1/day)	A	day	dCdt	date
30.01.20	15	9821	16419	0.469	482.885	13	1927	30.01.20
31.01.20	16	11948	18165	0.45	452.557	13	2044	30.01.20
1.02.20	17	14551	21823	0.414	394.184	14	2260	31.01.20
2.02.20	18	17389	26068	0.383	350.318	15	2495	1.02.20
3.02.20	19	20628	31257	0.354	316.693	16	2766	2.02.20
4.02.20	20	24553	38821	0.324	290.892	17	3146	3.02.20
5.02.20	21	28276	44479	0.308	279.637	18	3420	4.02.20
6.02.20	22	31439	46180	0.303	275.4	18	3496	4.02.20
7.02.20	23	34876	48078	0.297	268.19	18	3570	4.02.20
8.02.20	24	37552	48596	0.295	265.417	18	3588	4.02.20
9.02.20	25	40553	49822	0.291	256.799	19	3621	5.02.20
10.02.20	26	43099	50903	0.286	247.807	19	3643	5.02.20
11.02.20	27	44919	51372	0.284	243.22	19	3650	5.02.20
12.02.20	28	60326	71335	0.226	152.937	22	4027	8.02.20
13.02.20	29	64437	93484	0.198	137.983	24	4627	10.02.20
14.02.20	30	67100	100456	0.192	137.028	25	4826	11.02.20
15.02.20	31	69197	97849	0.194	137.789	25	4757	11.02.20
16.02.20	32	71329	94252	0.198	140.336	24	4672	10.02.20
17.02.20	33	73332	91467	0.202	143.931	24	4617	10.02.20
18.02.20	34	75198	89575	0.205	147.671	24	4589	10.02.20
19.02.20	35	75700	87525	0.208	153.464	24	4568	10.02.20
20.02.20	36	76676	86067	0.212	159.069	23	4560	9.02.20
21.02.20	37	77673	85090	0.214	163-862	23	4561	9.02.20
22.02.20	38	78651	84468	0.216	167.584	23	4564	9.02.20
23.02.20	39	79400	84039	0.217	170.598	23	4568	9.02.20
24.02.20	40	80088	**83756**	0.218	172.864	23	4573	9.02.20
25.02.20	41	80997	**83642**	0.219	173.897	23	4575	09.02.20

Table 2. Results of SIR simulations. After day 28, the method of data collection changes.

	Day	N	S_∞	R_∞	$R_{\infty,n}/R_{\infty,n-1}$	β	γ	β/γ	R^2
12.02.20	28	551513	473888	77625	1.429	2.897	2.689	1.077	0.988
13.02.20	29	1300538	1189616	110922	1.048	4.026	3.854	1.045	0.988
14.02.20	30	1434310	1318035	116275	0.925	4.157	3.988	1.042	0.990
15.02.20	31	1203132	1095609	107523	0.932	3.905	3.730	1.047	0.991
16.02.20	32	1002252	901990	100262	0.953	3.624	3.441	1.053	0.992
17.02.20	33	864976	769448	95528	0.969	3.387	3.198	1.059	0.993
18.02.20	34	774076	681530	92546	0.969	3.205	3.010	1.065	0.994
19.02.20	35	683791	594070	89722	0.978	2.998	2.798	1.071	0.994
20.02.20	36	619431	531645	87787	0.985	2.835	2.630	1.078	0.994
21.02.20	37	574963	488460	86503	0.990	2.712	2.504	1.083	0.994
22.02.20	38	544793	459126	85667	0.993	2.624	2.413	1.087	0.995
23.02.20	39	522591	437513	85078	0.993	2.557	2.343	1.091	0.995
24.02.20	40	506492	421823	84669	0.995	2.506	2.291	1.094	0.995
24.02.20	41	497719	413262	84456	0.998	2.477	2.262	1.095	0.996

Table 3. Estimated logistic model parameters for data until 25 Feb, 2020

	Estimate	SE	tStat	pValue
K	83643	1239.9	67.46	3.5853e-41
r	0.21882	0.0083521	26.199	6.2629e-26
A	173.89	29.73	5.8492	9.1682e-07

```
Number of observations: 41, Error degrees of freedom: 38
Root Mean Squared Error: 2.15e+03
R-Squared: 0.996, Adjusted R-Squared 0.995
F-statistic vs. zero model: 6.35e+03, p-value = 2.42e-51
```

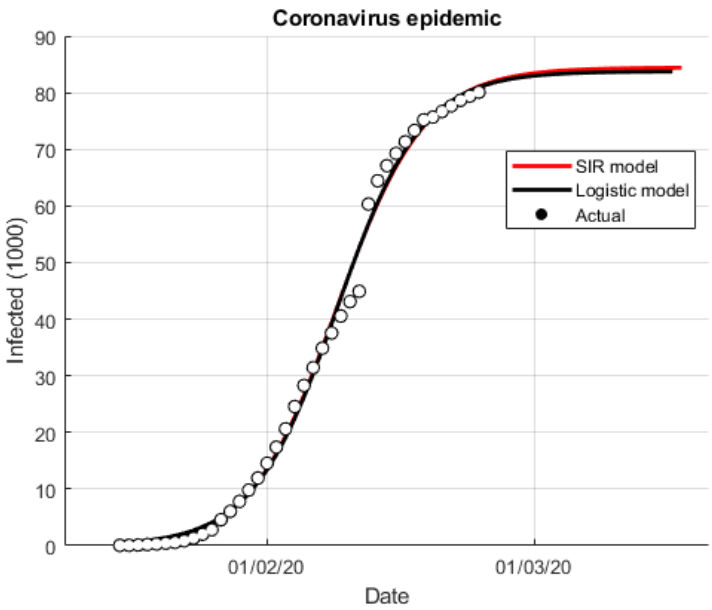

Figure 2. Predicted evaluation of coronavirus epidemic (data until 20 Feb, 2020)

In Figure 2, the time evaluation of the cases is shown, where we can see a good agreement between the models and the actual data. From Table 3, we see that the logistic regression model has a high coefficient of determination of 0.996, while the p-value (< 0.000) indicates that all the regression parameters are statistically significant.

In Tables 3 and 4, the iterated Shanks transformations for the predicted series of the final epidemic size are given. It appears that the predictions of the logistic model tend to the final size of 83231 cases, while the SIR model predictions converge to 83640 cases. Thus, the discrepancy is less than 0.5%.

Table 3. Iterated Shanks transformation for logistic model

day	K	$K(K)$	$K(K(K))$	$K(K(K(K)))$
35	87524			
36	86067	83101		
37	85090	83386	83236	
38	84469	83071	83372	83231
39	84039	90072	86718	
40	83576	83634		
41	83642			

Table 4. Iterated Shanks transformation for SIR model

day	R_∞	$R(R_\infty)$	$R(R(R_\infty))$	$R(R(R(R_\infty)))$
35	89722			
36	87787	83971		
37	86503	84107	84003	
38	85667	83673	83731	83640
39	85078	83740	83663	
40	84669	84225		
41	84456			

5 Short term forecasting

The models used are data-driven, so they are as reliable as data are. Namely, as can be seen from the graph in Figure 2 at the beginning, we have exponential growth. Then until 11 Feb, one can predict the final epidemic size of about 55000 cases. However, the collection of data changes and we have a jump of about 15000 new cases on 12. Feb. On 20 Feb we have another change in trend; the data begin to shows almost linear trend (See Fig 3). While the above models show that the epidemic is slow down, the linear trend predicts about 873 new cases per day (see Table 5).

Table 5. Short term forecasting with the logistic and linear model. The linear model predicts 873 new cases per day.

Date	Day	Actual	Logistic model	Error %	Linear model	Error %	Actaul cases/day	Predicted cases/day
20.02.20	35	75700	75890	2.111	75837	0.180		
21.02.20	36	76676	77298	2.337	76709	0.044	976	1408
22.02.20	37	77673	78468	2.267	77582	0.117	997	1170
23.02.20	38	78651	79434	2.004	78455	0.249	978	966
24.02.20	39	79400	80227	1.859	79328	0.090	749	793
25.02.20	40	80088	80876	1.644	80201	0.141	688	649
26.02.20	41	80997	81405	1.036	81074	0.095	909	529
27.02.20	42		81836		81947			431
28.02.20	43		82185		82820			349
29.02.20	44		82467		83693			282
1.03.20	45		82696		84566			229
2.03.20	46		82880		85439			184
3.03.20	47		83029		86312			149

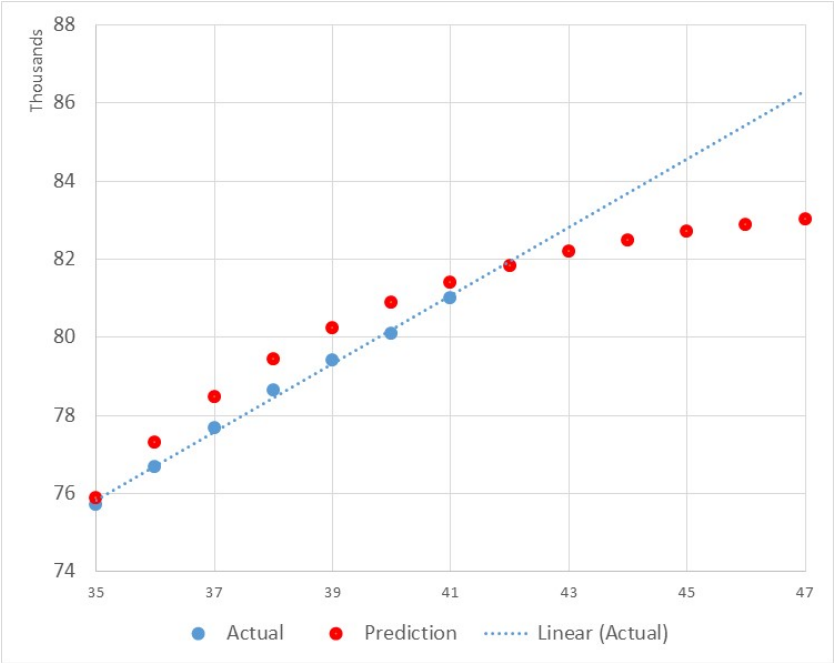

Figure 3. Short-term forecasting from 20 Feb 2020

6 Conclusion

On the basis of the available data, we can now predict that the final size of the coronavirus epidemic using the logistic model will be approximately 83700 (± 1300) cases and that the peak of the epidemic was on 9 Feb 2020. A more optimistic final size of 83300 cases is obtained using the Shanks transformation. Similar figures are obtained using the SIR model, where the predicted size of the epidemic is approximately 84500, and the Shanks transformation lowers this number to about 83700 cases. Naturally, the degree of accuracy of these estimates remains to be seen.

In conclusion, qualitatively, both models show that the epidemic is moderating, but recent data show a linear upward trend. The next few days will, therefore, indicate in which direction the epidemic is heading.

PS. Today it is more or less clear that the predictions of the article apply only to China. By February 20, 99% of the case was from China. The linear trend in data from Feb 20 onward meant a decreasing number of infected in China and increasing infected elsewhere in the world. In other words, in China, the epidemic is slowing down,

however, it is now developing elsewhere in the world. We note that the forecasting methods used in this article are inapplicable in the early stages of an epidemic.

References

Anastassopoulou, Cleo, Lucia Russo, Athanasios Tsakris, and Constantinos Siettos. 2020. "Data-Based Analysis, Modelling and Forecasting of the novel Coronavirus (2019-nCoV) outbreak." *medRxiv*:2020.02.11.20022186. doi: 10.1101/2020.02.11.20022186.

Bender, Carl M., and Steven A. Orszag. 1999. *Advanced mathematical methods for scientists and engineers I asymptotic methods and perturbation theory.* New York: Springer.

Brauer, Fred. 2019a. "Early estimates of epidemic final sizes." *Journal of Biological Dynamics* 13 (sup1):23-30. doi: 10.1080/17513758.2018.1469792.

Brauer, Fred. 2019b. "The Final Size of a Serious Epidemic." *Bulletin of mathematical biology* 81 (3):869-877. doi: 10.1007/s11538-018-00549-x.

Chowell G, Simonsen L, Viboud C, Kuang Y. 2014. "West Africa Approaching a Catastrophic Phase or is the 2014 Ebola Epidemic Slowing Down? Different Models Yield Different Answers for Liberia. ." *PLOS Currents Outbreaks.* doi: 10.1371/currents.outbreaks.b4690859d91684da963dc40e00f3da81.

Danby, J. M. A. 1985. *Computing applications to differential equations modelling in the physical and social sciences.* Reston, Va.: Reston Publishing Company.

Fisman D, Khoo E, Tuite A. . 2014. "Early Epidemic Dynamics of the West African 2014 Ebola Outbreak: Estimates Derived with a Simple Two-Parameter Model." *PLOS Currents Outbreaks.* . doi: 10.1371/currents.outbreaks.89c0d3783f36958d96ebbae97348d571.

Haberman, Richard. 1998. *Mathematical models mechanical vibrations, population dynamics, and traffic flow an introduction to applied mathematics.* Unabridged republication ed, *Classics in applied mathematics.* Philadelphia: SIAM.

Hethcote, Herbert W. 2000. "The Mathematics of Infectious Diseases." *SIAM Review* 42 (4):599-653. doi: 10.1137/S0036144500371907.

Miller, Joel C. 2012. "A note on the derivation of epidemic final sizes." *Bulletin of mathematical biology* 74 (9):2125-2141. doi: 10.1007/s11538-012-9749-6.

Murray, James Dickson. 2002. *Mathematical biology*. 3rd ed, *Interdisciplinary applied mathematics*. New York: Springer.

Nesteruk, Igor. 2020. "Statistics based predictions of coronavirus 2019-nCoV spreading in mainland China." *medRxiv*:2020.02.12.20021931. doi: 10.1101/2020.02.12.20021931.

Pell, Bruce, Yang Kuang, Cecile Viboud, and Gerardo Chowell. 2018. "Using phenomenological models for forecasting the 2015 Ebola challenge." *Epidemics* 22:62-70. doi: https://doi.org/10.1016/j.epidem.2016.11.002.

Roosa, K.;, Y.; Lee, R.; Luo, A.; Kirpich, R.; Rothenberg, J.M.; Hyman, P.; Yan, and G. Chowell. 2020. "Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China." *Journal of Clinical Medicine* 9 (2):596.

Shanks, Daniel. 1955. "Non-linear Transformations of Divergent and Slowly Convergent Sequences." *Journal of Mathematics and Physics* 34 (1-4):1-42. doi: 10.1002/sapm19553411.

Wu, Joseph T., Kathy Leung, and Gabriel M. Leung. 2020. "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study." *The Lancet*. doi: https://doi.org/10.1016/S0140-6736(20)30260-9.

Xiong, Hao, and Huili Yan. 2020. "Simulating the infected population and spread trend of 2019-nCov under different policy by EIR model." *medRxiv*:2020.02.10.20021519. doi: 10.1101/2020.02.10.20021519.