Estimation of the final size of the coronavirus epidemic by the logistic model

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

(Feb 2020)

Abstract

In the note, the logistic growth regression model is used for the estimation of the final size and its peak time of the coronavirus epidemic. Based on available data estimation the final size will be about 90 000 cases and the peak time was on 10 Feb 2020.

1 Introduction

One of the common questions in an epidemic is its final size and its peak time. For answering this question various models are used: analytical [1-4], stochastic [5], phenomenological [6, 7]

Xiong and Yan [8], using EIR model, for data til 7. Feb 2020 predicts that the final size of the epidemy size would be 115 022 infections. For data for the period between 16 Jan 2020 to 8 Feb 2020 Nesteruk [9], using the SIR model, prediction was about 65 000. For data til 10 Feb 2020 Anastassopoulou et all [10], using the SIRD model, forecast 140 000 (scenario I) to 1 million (Scenario 2) infections on 29 Feb 2020.

In this note, we will try to estimate a final epidemic size and its peak by the phenomenological model, i.e., with the logistic growth regression model [7, 11]. The estimation is made in two steps: first, data are sequentially fitted by logistic model and second, the estimated finale sizes are fitted by the Weibull function.

2 Logistic grow model

The logistic growth model originates in population dynamics [12]. 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. So, if C is the number of cases, and t is the time, then the models read

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

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

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

where $A = \frac{K - C_0}{C_0}$ The growth rate $\frac{dC}{dt}$ reaches its maximum when $\frac{d^2C}{dt^2} = 0$. From this condition, we obtain that the growth rate peak occurs in time time

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

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

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

3 Logistic regression model

By the logistic regression model, we estimate the parameters of the logistic model. Using (2) we can write the regression model in the form

$$C = \frac{b_1}{1 + b_2\exp(-b_3 t)} \tag{5}$$

where unknown parameters b_1, b_2, b_3 which are to be estimated from the data. In the logistic model, the parameters are constants while the regression model parameters depend on available data. So if $C_1, C_2, \ldots, C_n$ are the cases in times $t_1, t_2, \ldots, t_n$ then $b_i^{(n)} = b_i^{(n)}(C_1, C_2, \ldots, C_n)$ ($i = 1,2,3$), or in other words, parameters are time-dependent

$$b_i = b_i(t) \quad (i = 1,2,3) \tag{6}$$

Now, if we assume that the epidemy has a final size then $b_i = b_i(t)$ should converge to some finite values.

Alternatively, we can fit the sequence $b_i^{(k)}$, $k = n_0, \ldots, n$ by Weibull function (or any other suitable function)

$$b_1 = K\left[1 - \exp\left(-k\left(t - t_0\right)^p\right)\right] \tag{7}$$

we obtain a final estimation of K, i.e., the epidemy size. k, p and t_0 are regression parameters.

4 Data

For regression analysis, we use two data sets.. For the period from 16 Jan to 21 Jan 2020 we use the data from the published graph[1] and for the period from 22 Jan 2020 to 19 Feb 2020 we use data from *worldmeters*[2]. The discrepancy between data set for a period of 22 Jan to 26 Jan is between 1% to 6%.

5 Results

For practical calculation, we use MATLAB's functions `lsqcurvefit` and `fitnlm` (see Apendix).

Based on available data (til 18 Feb 2020), the predicted final size of epidemy is about 89600 cases (see Table 1, see Fig 1, 2) and its peak was on 10 Feb 2020. The regression has a high coefficient of determination 0.993 while p-value (< 0.000) indicates that all parameters are statistically significant. With the new data, these estimates will change reliably within days.

Table 1. Estimate logistic model parameters for (data till 17.Feb 2020).

Estimate	SE	tStat	pValue	
b1	**89577**	3916.2	22.874	5.6716e-21
b2	0.20493	0.011318	18.107	4.8542e-18
b3	147.66	28.03	5.2681	9.9549e-06

```
Number of observations: 34, Error degrees of freedom: 31
Root Mean Squared Error: 2.24e+03
R-Squared: 0.993,  Adjusted R-Squared 0.993
F-statistic vs. zero model: 2.97e+03, p-value = 3.34e-38
```

[1] https://i.redd.it/f4nukz4ou9d41.png

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

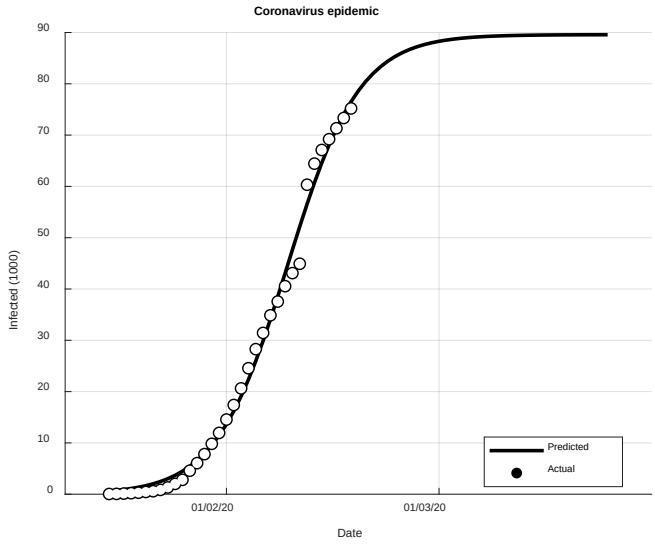

Figure 1. Predicted evaluation of the coronavirus epidemic (data till 19.Feb 2020). Regression data are in Table 4.

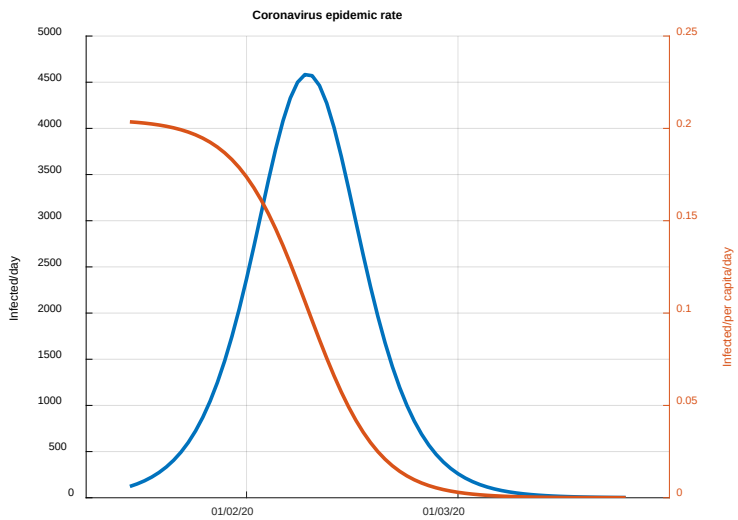

Figure 2. Predicted evaluation of the coronavirus epidemic rate (data till 19.Feb 2020)

The trend of model parameters is given in Table 2 while the graph of evaluation of the final size of the epidemy is shown as a bar graph in Fig 3. The parameters of the Weibull model (8) with $p = 1$, $t_0 = 0$, of evaluation of the size of epidemy, for data until 12. Feb 2020, are given in Table 3. We see that the final size of epidemy would be about 51000 infections and the epidemy would reach its peak on 5 Feb 2020. However, a new way of data collection distorts this prediction. By using the Weibull

model, we obtain the final size of the epidemy of about 117300 infections (see Table 4). We stress that this prediction should be taken with care because there is not yet enough data to use the Weibull model for the final prediction of epidemy size. Namely, from Table 4, we see that R^2 is relatively high (0.91) but regression coefficients for k have p-value > 0.1 so it is statistically insignificant. Similar observation also applies to the final prediction of the epidemy peak time (Fig 4), where Weibull regression predicts peak time at day 37, i.e., on 23 Feb 2020. In this regression three of regression coefficient are insignificant with p-val > 0.85.

Table 2. 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 (cases/day)	date
30.jan.20	15	9821	16419	0.469	482.885	13	1927	30.jan.20
31.jan.20	16	11948	18165	0.450	452.557	13	2044	30.jan.20
1.feb.20	17	14551	21823	0.414	394.184	14	2260	31.jan.20
2.feb.20	18	17389	26068	0.383	350.318	15	2495	1.feb.20
3.feb.20	19	20628	31257	0.354	316.693	16	2766	2.feb.20
4.feb.20	20	24553	38821	0.324	290.892	17	3146	3.feb.20
5.feb.20	21	28276	44479	0.308	279.637	18	3420	4.feb.20
6.feb.20	22	31439	46180	0.303	275.4	18	3496	4.feb.20
7.feb.20	23	34876	48078	0.297	268.19	18	3570	4.feb.20
8.feb.20	24	37552	48596	0.295	265.417	18	3588	4.feb.20
9.feb.20	25	40553	49822	0.291	256.799	19	3621	5.feb.20
10.feb.20	26	43099	50903	0.286	247.807	19	3643	5.feb.20
11.feb.20	27	44919	51372	0.284	243.22	19	3650	5.feb.20
12.feb.20	28	60326	71335	0.226	152.937	22	4027	8.feb.20
13.feb.20	29	64437	93484	0.198	137.983	24	4627	10.feb.20
14.feb.20	30	67100	100456	0.192	137.028	25	4826	11.feb.20
15.feb.20	31	69197	97849	0.194	137.789	25	4757	11.feb.20
16.feb.20	32	71329	94252	0.198	140.336	24	4672	10.feb.20
17.feb.20	33	73332	91467	0.202	143.931	24	4617	10.feb.20
18.feb.20	34	75198	89575	0.205	147.671	24	4589	10.feb.20

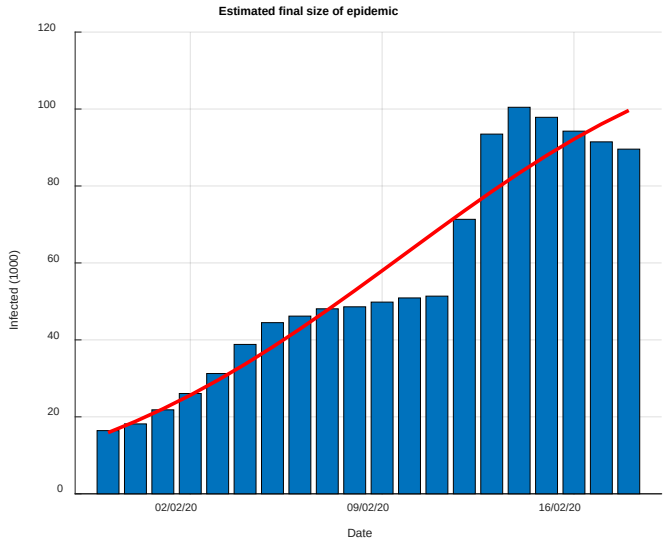

Figure 3. Evaluation of estimated final size of coronavirus epidemic (bars). Weibull regression data (red line) are given in Table 4.

Table 3. The estimated final size of epidemy from data till 12.Feb. 2020

	Estimate	SE	tStat	pValue
b1	**58121**	3746.8	15.512	8.424e-08
b2	0.18721	0.039055	4.7935	0.00098261
b3	12.085	7.1289	1.6952	0.12428

```
Number of observations: 12, Error degrees of freedom: 9
Root Mean Squared Error: 2.11e+03
R-Squared: 0.975,  Adjusted R-Squared 0.97
F-statistic vs. zero model: 1.53e+03, p-value = 1.7e-12
```

Table 4. The estimated final size of the epidemy by Weibull model (7) from data till 19.Feb. 2020 (see Fig 3)

	Estimate	SE	tStat	pValue
b1	**1.173e+05**	9.8924e-06	1.1858e+10	1.7123e-171
b2	4.2462e-07	5.8668e-07	0.72377	0.47851
b3	-8.009	0.049584	-161.53	6.5282e-30
b4	4.1226	0.37745	10.922	2.2606e-09

```
Number of observations: 20, Error degrees of freedom: 18
Root Mean Squared Error: 8.67e+03
R-Squared: 0.913,  Adjusted R-Squared 0.908
F-statistic vs. zero model: 520, p-value = 1.2e-15
```

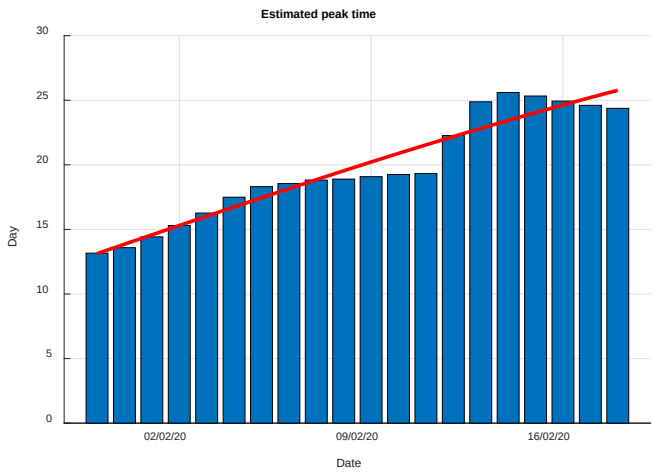

Figure 4. Evaluation of estimated peak time of coronavirus epidemic (bars). Weibull regression data (red line) (Table 5).

Table 5. The estimated peak time size of the epidemy by Weibull model (7) from data till 19.Feb. 2020 (see Fig 5)

Estimated Coefficients:

	Estimate	SE	tStat	pValue
b1	**37.545**	102	0.3681	0.71762
b2	0.00092784	0.046479	0.019963	0.98432
b3	-13.031	172.84	-0.075395	0.94084
b4	1.8621	12.228	0.15228	0.88087

```
Number of observations: 20, Error degrees of freedom: 16
Root Mean Squared Error: 1.22
R-Squared: 0.926,  Adjusted R-Squared 0.912
F-statistic vs. zero model: 1.35e+03, p-value = 5.15e-20
Peak date 23-Feb-2020
```

5 Conclusion

On the base of available data, one can now predict that the final size of coronavirus epidemy using the logistic model will be around 90 000 cases. The peak of the epedimiy was on 10 Feb 2020.

Less reliable is the following prediction with Weibul function: the final number of infections will be about 115 000, and that peak will be at about 37 days, i.e., on 23. Feb 2020.

Appendix. MATLAB program.

```
%FITVIRUS Estimation of coronavirus epidemy size by logistic model

close all

warning('off')

% obtain data
[sampleC,date0] = getData();


% form time
samplaTime = 0:1:length(sampleC)-1;
time = 0:1:2*length(samplaTime);

% form date
samplaDate = date0+samplaTime;
date = date0+time;

% initial guess
b0 = [max(sampleC) 1 max(sampleC)]';

% parameters evaluation
bb = NaN(3,length(sampleC));
tpeak = NaN(length(sampleC),1);
dCpeak = NaN(length(sampleC),1);

%initial data set
n0 = 15;

% loop over periods
opts = optimoptions('lsqcurvefit','Display','off');
for n = n0:length(sampleC)
    [b,resnorm,~,exitflag,output] = lsqcurvefit(@fun,b0,...
        samplaTime(1:n),sampleC(1:n),[],[],opts);
    bb(:,n) = b;
    % peak time and growth rate
    tpeak(n) = log(b(3))/b(2);
    dCpeak(n) = dfun(b,tpeak(n));
end

% final fit
fprintf('Regression parameters for complet data set\n')
mdl = fitnlm(samplaTime(1:n),sampleC(1:n),@fun,b0)

% print parameters evaluation
fprintf('\nEvaluation of model parameters\n')
fprintf('%4s %12s %10s %10s %10s %12s
%12s\n','day','K','r','A','t_peak','dC_peak','date_peak')
fprintf('%4s %12s %10s %10s %10s %12s\n','
','(cases)','(1/day)','','(day)','(cases/day)')
for n = n0:length(sampleC)
    fprintf('%4d %12d %10.3f %10.3f %10d %12d
%12s\n',n,fix(bb(1,n)),bb(2:3,n),...
        fix(tpeak(n)),fix(dCpeak(n)),datestr(ceil(tpeak(n))+date0))
end

% fit estimated final sizes
fprintf('\nWeibull regression of predicted final sizes\n')
```

```
% Initial guess
f0 = [sampleC(end) 0.13 12 1]'
n = length(sampleC);
mdl2 = fitnlm(samplaTime(n0:n),bb(1,n0:n),@fun1,f0)
f = mdl2.Coefficients.Estimate;

% fit estimated peak time
fprintf('\nWeibull regression of predicted peak time\n')
% Initial guess
ff0 = [tpeak(end) 0.13 12 1]'
n = length(sampleC);
mdl2 = fitnlm(samplaTime(n0:n),tpeak(n0:n),@fun1,ff0)
ff = mdl2.Coefficients.Estimate;
fprintf('Peak date %s\n',datestr(ceil(ff(1))+date0))

warning('on')

% plot evaluation of peak time
figure
hold on
bar(samplaDate(n0:n),tpeak(n0:n)); %,'LineWidth',2)
datetick('x',20,'keeplimits')
plot(date(n0:n),fun1(ff,time(n0:n)),'r','LineWidth',2)
datetick('x',20,'keeplimits')
ylabel('Day')
xlabel('Date')
title('Estimated peak time')
grid on
hold off

% plot evaluation of final size
figure
hold on
bar(samplaDate(n0:n),bb(1,n0:n)/1000); %,'LineWidth',2)
datetick('x',20,'keeplimits')
plot(date(n0:n),fun1(f,time(n0:n))/1000,'r','LineWidth',2)
datetick('x',20,'keeplimits')
ylabel('Infected (1000)')
xlabel('Date')
title('Estimated final size of epidemic')
grid on
hold off

% plot final prediction of epedimy evaluation
figure
hold on
plot(date,fun(b,time)/1000,'k','LineWidth',2)
scatter(samplaDate,sampleC/1000,50,'k','filled')
h = scatter(samplaDate,sampleC/1000,30,'w','filled');
h.Annotation.LegendInformation.IconDisplayStyle = 'off';
% add tick marks
datetick('x',20,'keeplimits')
% add axes labels
ylabel('Infected (1000)')
xlabel('Date')
% add legend
legend('Predicted','Actual','Location','best')
%add title
title('Coronavirus epidemic')
% add grid
grid on
```

```
hold off

% plot growth rate of epidemic
figure
hold on
plot(date,dfun(b,time),'LineWidth',2)
ylabel('Infected/day')
yyaxis right
plot(date,dfun(b,time)./fun(b,time),'LineWidth',2)
ylabel('Infected/per capita/day')
datetick('x',20,'keeplimits')
title('Coronavirus epidemic rate')
grid on
hold off

function y = fun(b,t)
% Logistic model
    y = b(1)./(1 + b(3)*exp(-b(2)*t));
end

function y = dfun(b,t)
% Grow rate
    y = b(1)*b(2)*b(3)*exp(-b(2)*t)./(1 + b(3)*exp(-b(2)*t)).^2;
end

function y = fun1(b,t)
% Weibull model. Note b(3) !!!
    y = b(1)*(1 - exp(-b(2)*(t-b(3)).^b(4)));
end

function [C,date0] = getData()
%GETDATA Coronavirus data
%   data from 16 Jan to 21 Jan https://i.redd.it/f4nukz4ou9d41.png
%   data from from 22 Jan 2020 to 18 Feb 2020 are from
%   https://www.worldometers.info/coronavirus/
date0=datenum('2020/01/16'); % start date
C = [
45
62
121
198
291
440
580
845
1317
2015
2800
4581
6058
7813
9821
11948
14551
17389
20628
24553
28276
31439
34876
37552
```

```
     40553
     43099
     44919
     60326
     64437
     67100
     69197
     71329
     73332
     75198
%<-------------- add new data here
     ]';
     end
```

References

[1] D.A. Barbarossa MV, Kiss G, Nakata Y, Röst G, Vizi Z Transmission Dynamics and Final Epidemic Size of Ebola Virus Disease Outbreaks with Varying Interventions, PLoS ONE, 10 (2015).

[2] F. Brauer, The Final Size of a Serious Epidemic, Bull Math Biol, 81 (2019) 869-877.

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

[4] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, The Lancet, (2020).

[5] J.C. Miller, A note on the derivation of epidemic final sizes, Bull Math Biol, 74 (2012) 2125-2141.

[6] K.E. Fisman D, Tuite A. , Early Epidemic Dynamics of the West African 2014 Ebola Outbreak: Estimates Derived with a Simple Two-Parameter Model., PLOS Currents Outbreaks. , (2014).

[7] B. Pell, Y. Kuang, C. Viboud, G. Chowell, Using phenomenological models for forecasting the 2015 Ebola challenge, Epidemics, 22 (2018) 62-70.

[8] H. Xiong, H. Yan, Simulating the infected population and spread trend of 2019-nCov under different policy by EIR model, medRxiv, (2020) 2020.2002.2010.20021519.

[9] I. Nesteruk, Statistics based predictions of coronavirus 2019-nCoV spreading in mainland China, medRxiv, (2020) 2020.2002.2012.20021931.

[10] C. Anastassopoulou, L. Russo, A. Tsakris, C. Siettos, Data-Based Analysis, Modelling and Forecasting of the novel Coronavirus (2019-nCoV) outbreak, medRxiv, (2020) 2020.2002.2011.20022186.
[11] S.L. Chowell G, Viboud C, Kuang Y., West Africa Approaching a Catastrophic Phase or is the 2014 Ebola Epidemic Slowing Down? Different Models Yield Different Answers for Liberia. , PLOS Currents Outbreaks., (2014).
[12] R. Haberman, Mathematical models mechanical vibrations, population dynamics, and traffic flow an introduction to applied mathematics, Unabridged republication ed., SIAM, Philadelphia, 1998.