LogEpi: Logistic curves applied to epidemiology

Beatriz Cuyabano and Gabriel Rovere

March 29, 2020

This vignette offers a brief explanation of how the logistic curve is used to model epidemiological events, and provides a set of examples.

A logistic curve can be used to fit a counting process that starts at zero and ends at a determined number. Such curve has been widely used to describe the growth of a population, the growth of cases in a disease outbreak, or the growth of deaths in a such outbreak. The general logistic function is described as:

$$y \;=\; f(x \mid a, b, c) \;=\; \frac{a}{1 + e^{-(x-b)/c}}. \tag{1}$$

In equation (1), the variable x represents the days passed after the first occurrence of the event to be described, and a, b, and c are the parameters of this event's process. To simplify the description, $x = 1$ at the day of the first occurrence.

- a is the total number of occurrences at the stabilization of the process,
- b is the day in which the maximum new occurences of the event will happen,
- c is the speed of infection.

The parameters are obtained by minimizing the mean square error of the model:

$$MSE \;=\; MSE(a, b, c \mid \boldsymbol{x}, \boldsymbol{y}) \;=\; \sum_{i=1}^{n} \left[y_i - f(x_i \mid a, b, c)\right]^2 \;=\; \sum_{i=1}^{n} \left[y_i - \frac{a}{1 + e^{-(x_i-b)/c}}\right]^2, \tag{2}$$

$$(\hat{a}, \hat{b}, \hat{c}) = argmin \left[MSE(a, b, c \mid \boldsymbol{x}, \boldsymbol{y})\right]. \tag{3}$$

It is important to remark that this curve does not take into account any demographic or social information. It merely adjusts to the observed numbers provided. Before the speed of infection stabilizes, it is necessary to make some projections to model the entire outbreak. Once the speed of infection is stabilized, the curves are reliable without projections. This will be detailed with examples in this document. The Covid-19 data in use for the examples is downloaded directly into R from the European Centre for Disease Prevention and Control (`https://www.ecdc.europa.eu`) webpage.

To start the analysis, call the `LogEpi` library, and load the Covid-19 dataset from the European Centre for Disease Prevention and Control. The examples and analysis displayed next are on the data of March 29, 2020.

```
# load package
library(LogEpi)

# load data from the European Centre for Disease Prevention and Control
info.url <- paste("https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-geographic-disbtribution-world
wide-",format(Sys.time(),"%Y-%m-%d"),".xlsx",sep="")

# save into a temporary file
GET(info.url,authenticate(":",":",type="ntlm"),write_disk(tf <- tempfile(fileext=".xlsx")))

# convert to data.frame and some minor adjustments to variables names
info <- as.data.frame(read_excel(tf))[,-1]
names(info)[6] <- "Countries.and.territories"

# check loaded data
# this data base contains the reported cases and deaths daily, in each country and territory
head(info)

#   day month year cases deaths Countries.and.territories geoId countryterritoryCode popData2018
# 1  29     3 2020    15      1               Afghanistan    AF                  AFG    37172386
# 2  28     3 2020    16      1               Afghanistan    AF                  AFG    37172386
# 3  27     3 2020     0      0               Afghanistan    AF                  AFG    37172386
# 4  26     3 2020    33      0               Afghanistan    AF                  AFG    37172386
# 5  25     3 2020     2      0               Afghanistan    AF                  AFG    37172386
# 6  24     3 2020     6      1               Afghanistan    AF                  AFG    37172386

# display countries sorted decreasingly by the total number of cases
sort.totals <- aggregate(cases ˜ Countries.and.territories,data=info,FUN=sum)
sort.totals <- sort.totals[sort(sort.totals$cases,decreasing=TRUE,index.return=TRUE)$ix,]
rownames(sort.totals) <- 1:nrow(sort.totals)
head(sort.totals)

#    Countries.and.territories  cases
# 1   United_States_of_America 124665
# 2                      Italy  92472
# 3                      China  82342
# 4                      Spain  72248
# 5                    Germany  52547
# 6                     France  37575
# 7                       Iran  35408
# 8             United_Kingdom  17089
# 9                Switzerland  13152
# 10               Netherlands   9762
# 11               South_Korea   9583
# 12                   Belgium   9134
# 13                   Austria   8291
# 14                    Turkey   7402
# 15                    Canada   5386
# 16                  Portugal   5170
# 17                    Brazil   3904
# 18                    Norway   3845
# 19                 Australia   3809
# 20                    Israel   3619
```

Now, the functions from `LogEpi` library will be used to extract the data on separate countries, and to generate the logistic curves to describe the growth of cases and deaths on the Covid-19 outbreak.

```
# create table with cumulative data in China
data <- mkEpiTables("China")
head(data,20)

#          China cases deaths
# 1  2019-12-31    27      0
# 2  2020-01-01    27      0
# 3  2020-01-02    27      0
# 4  2020-01-03    44      0
# 5  2020-01-04    44      0
# 6  2020-01-05    59      0
# 7  2020-01-06    59      0
# 8  2020-01-07    59      0
# 9  2020-01-08    59      0
# 10 2020-01-09    59      0
# 11 2020-01-10    59      0
# 12 2020-01-11    59      1
# 13 2020-01-12    59      1
# 14 2020-01-13    59      1
# 15 2020-01-14    59      1
# 16 2020-01-15    59      2
# 17 2020-01-16    59      2
# 18 2020-01-17    63      2
# 19 2020-01-18    80      2
# 20 2020-01-19   216      3

# run analysis with the logistic curves
fitEpi <- mkEpiCurves(data)
```

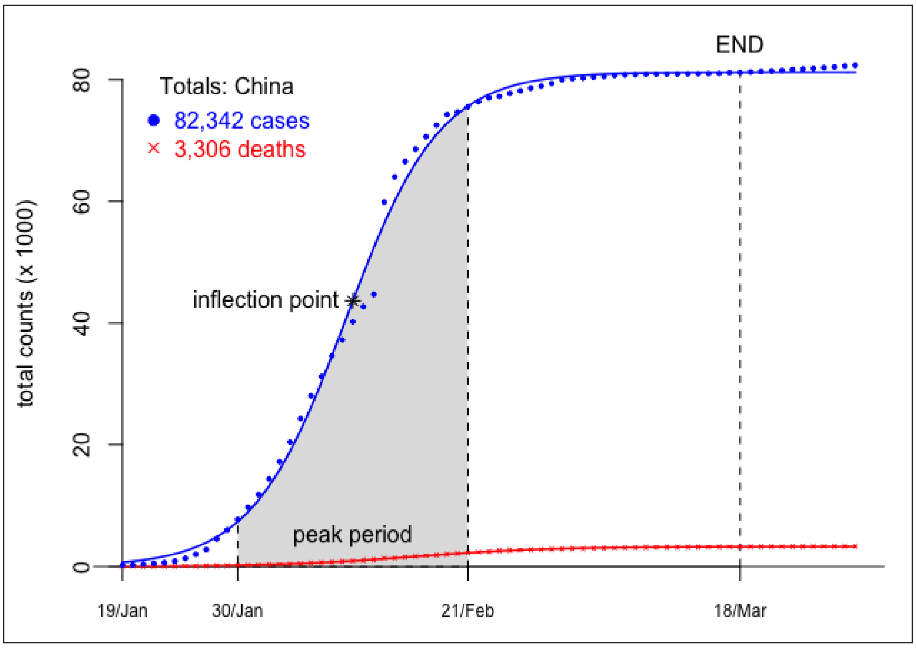

Figure 1: Plot of data and fitted curves in China.

```
# check which was considered as day 1 to model outbreak
fitEpi$day1

# [1] "2020-01-19"

# call the parameters
fitEpi$parameters

#               a        b        c
# cases 81209.364 22.32903 4.474012
# death  3254.264 19.97771 6.425636
```

Parameter a indicates the expected number of cases and deaths at the date considered the end of the outbreak (81,210 and 3,255 respectively).

Parameter b indicates how many days after day 1 (2020-01-19) the outbreak will reach its peak for cases and deaths. Rounding these numbers to 22 and 20, the 2020-02-10 is the date for the peak of cases and 2020-02-08 is the date for the peak of deaths. For the cases, this date is the inflection point indicated in the plot.

Parameter c indicates the speed of infection and of occurrence of deaths. $1/c$ is the mean daily increase rate until the peak. In this example, $1/c \approx 1/4.474 \approx 0.224$, meaning that on average, until the date 2020-02-10, the number of new daily cases was 22.4% of the total number of cases in the previous day.

```
# this example will call the table directly in the function to generate the curves
fitEpi <- mkEpiCurves(mkEpiTables("Italy"))
```

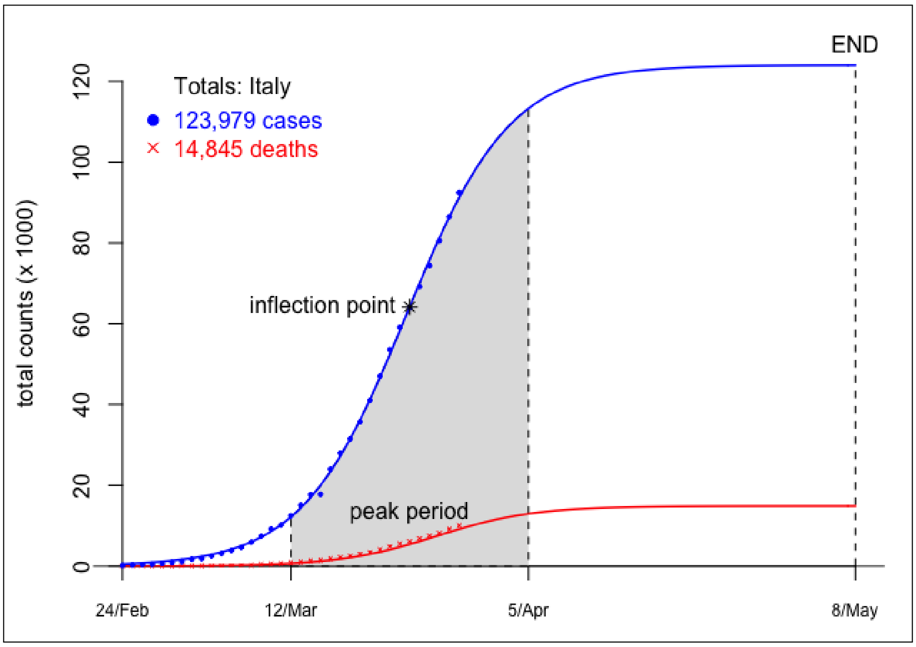

Figure 2: Plot of data and fitted curves in Italy.

For countries in which the number of cases is still low, a projection of cases on the coming days is needed to allow the fit of the logistic curve. The next example shows a projection of 7 days.

```
# this example will call the table directly in the function to generate the curves
fitEpi <- mkEpiCurves(mkEpiTables("New_Zealand"),project=7)
```

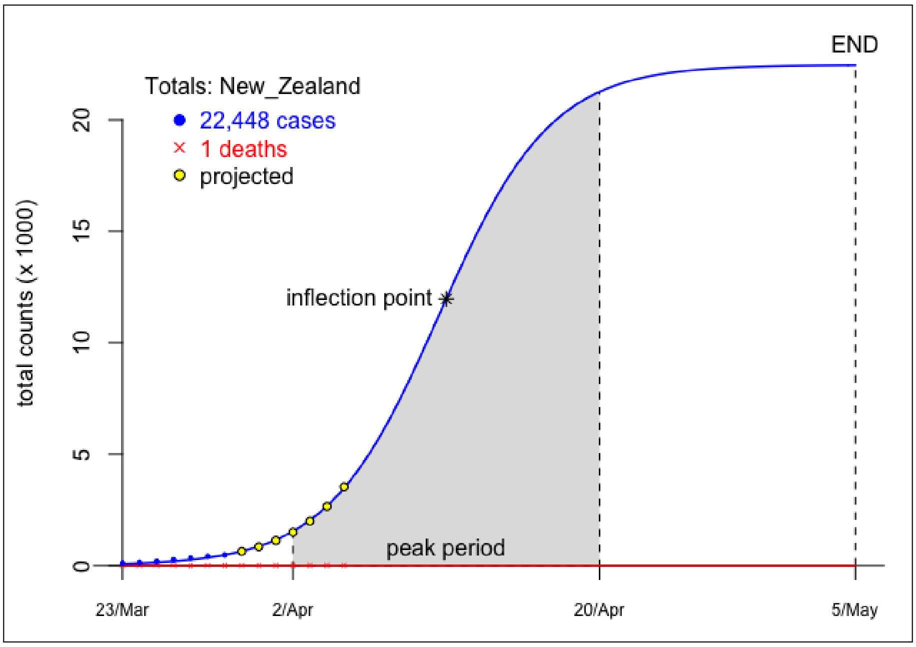

Figure 3: Plot of data and fitted curves in New Zealand.

`LogEpi` is in its initial version. Please help us improve our work by reporting bugs, errors and suggestions to bia.cdc@gmail.com.