Nonlinear regression of COVID19 infected cases.

- Posted by Pablo Gutierrez on April 12, 2020 at 11:23am
- View Blog

In 1927, W. O. Kermack y A. G. McKendrick described the first mathematical model for infectious diseases using a set of differential eq
SIR because of the three states one individual can have.
These states are:

- Susceptible: The individuals that can be infected by the disease
- Infected: The individuals that have been infected and suffer the disease.
- Recovered: The individuals that recovered from the disease and have become immune.

The equations that represent these states are as follows:

1. Variation with time of the susceptible individuals to be infected will depend inversely on a transmission factor β and the susceptit
$$\frac{dS(t)}{dt} = -\beta S(t) I(t)$$
2. Variation of those infected will depend on the number of people who are still susceptible of being infected, minus the number of
$$\frac{dI(t)}{dt} = -\beta S(t) I(t) - \alpha I(t)$$
recovered and are therefore immune.
3. The variation of recovered ones depends directly on the number of infected multiplied by α, a factor that determines the time tha
$$\frac{dR(t)}{dt} = \alpha I(t)$$
is:

The boundary conditions are:

- Population must always remain constant$N(t) = S(t) + I(t) + R(t)$At t=0

$$I(t) = 0 R(t) = 0$$

The analytical solution of this system can be found in different articles, for example here: arXiv:1403.2160

Instead of that, I will focus in equation (2) to note that it is a Bernoulli equation of the form

$$\psi(x) = -\beta S(x), \phi(x) = -\alpha I(x)$$

The solution for this Bernoulli differential equation is the **logistic** function, which most general form is this:

$$N(t) = a + \left(\frac{b}{1 + e^{-\frac{t-c}{d}}}\right)$$

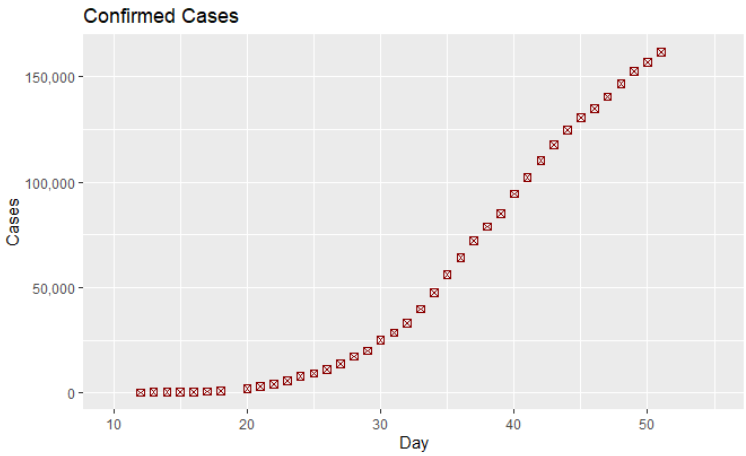
In the epidemiologic context, this logistic function represents the accumulative number of infected people as a function of time.

Using this model, it's possible to fit it to the real data, to obtain the values for the variables, the way to do it consists in minimizing the re

$$RSS(\beta) = \sum_{i}^{n} (y_i - f(x_i, \beta))^2$$

Because the function to be fitted is not linear, the method to minimize de loss function must be regressions. To do this regression, I used the NLS package for R, which implements the Gauss-Newton algorithm.

The data corresponds to the number of infected people in Spain as a function of time provided by the Ministry of Health.



This graph represents the c

How to execute the regression using R.

1. Load the CSV with data using read_csv

descarga <- read_csv("serie_historica_acumulados.csv",col_types = colsFallecidos = col_double(), Fecha = col_date(format = "%d/%n /%Y"), Hospitalizados = col_double(), Recuperados = col_double(), UCI = col_double(), X8 = col_skip()))

2. Group the data by date and sum all regions

agregados_por_fecha<-descarga %>% group_by(Fecha) %>% summarize(Fallecidos=sum(Fallecidos), Casos=sum(Casos), Hospitali

3. Create a sequence to use it as a time scale

s<-seq(1:length(tabla_absolutos$Fecha))

tabla_absolutos["dia"] <- s

4. Use nls to fit the curve. To have a good fit, it is necessary to provide initial data compatible with the data. This need to be made r

logis.m1 <- nls(Casos ~ logis(dia, a, b, c,d), data = agregados_por_fecha, start = list(a = 0, b = 180000, c = 40, d=5))

5. Use summary to retrieve the details of the regression.

summary(logis.m1)

Formula: Casos ~ logis(dia, a, b, c, d)

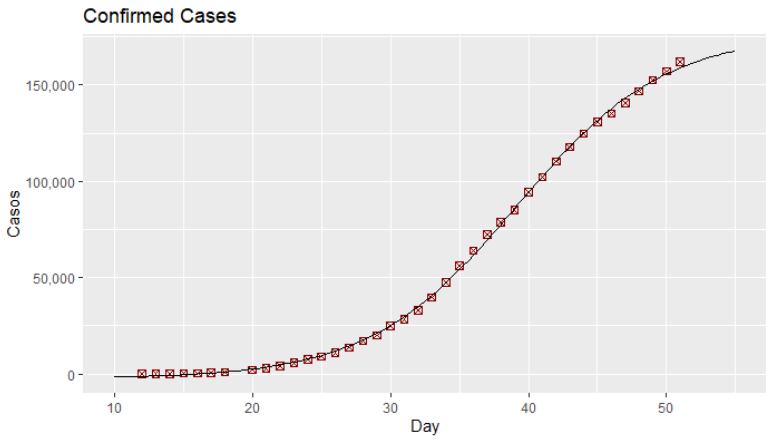
Parameters:

b 1.788e+05 2.111e+03 84.706 < 2e-16 ***

c 3.914e+01 1.317e-01 297.217 < 2e-16 ***

d 5.362e+00 1.033e-01 51.920 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



This graph represents the data and the

Conclusions:

- The regression found the values for the variable that are compatible withe the data.
- The inflexion point occurred on day 39 (march29)
- The maximum number of infected people will be 180.000 people
- The number of infected will grow until May 15th.

Most Popular Content on DSC

To not miss this type of content in the future, subscribe to our newsletter.

- Book: Statistics -- New Foundations, Toolbox, and Machine Learning Recipes
- Book: Classification and Regression In a Weekend - With Python
- Book: Applied Stochastic Processes
- Long-range Correlations in Time Series: Modeling, Testing, Case Study
- How to Automatically Determine the Number of Clusters in your Data
- New Machine Learning Cheat Sheet | Old one
- Confidence Intervals Without Pain - With Resampling
- Advanced Machine Learning with Basic Excel
- New Perspectives on Statistical Distributions and Deep Learning
- Fascinating New Results in the Theory of Randomness
- Fast Combinatorial Feature Selection

Other popular resources

- Comprehensive Repository of Data Science and ML Resources

Views: 3151

- < Previous Post

Comment

You need to be a member of Data Science Central to add comments!

Join Data Science Central

 Comment by Pablo Gutierrez on April 16, 2020 at 1:15am

Hi Jason.

The data is for cumulative cases. Of course most of these people will recover after a while, but the analysis is focused on the inf

Regards

 Comment by Jason Chia Kim Leng on April 15, 2020 at 7:34pm

This seems to model a scenario where the number of infected COVID-19 cases will plateau eventually instead of decline back d that life will never return back to normal...unless the graphical plot's y axis represents total cumulative case count since day of in account the number of cases that have recovered, but there is minimal description and/or annotation of the plot to firm up a fixed

 Comment by Pablo Gutierrez on April 13, 2020 at 10:16pm

Thanks Peter. I will check it out. Regards

 Comment by Peter Cotton on April 13, 2020 at 9:02am

Nice exposition. Thanks. Of course there are some issues with assuming a representative agent. See this post for some discuss and simulations using pandemic on PyPI. I'll follow up with a blog article here.

 Comment by Habib on April 12, 2020 at 7:30pm

Could you please share the R codes and the data files in English to replicate the results at:

habibnawazbnu@gmail.com

RSS

RESOURCES

- Subscribe to DSC Newsletter
- Free Books
- Forum Discussions
- Cheat Sheets
- Jobs
- Search DSC
- DSC on Twitter
- DSC on Facebook

VIDEOS

DSC Webinar Series: Accelerating AI Adoption with Machine Learning Operations (MLOps)

Added by Tim Matteson 0 Comments 2 Likes

DSC Webinar Series: 500 Petabytes of Data to Understand the Universe Better

Added by Tim Matteson 0 Comments 1 Like

DSC Webinar Series: Mathematical Optimization Modeling: Learn the Basics

Added by Tim Matteson 1 Comment 1 Like

- Add Videos
- View All