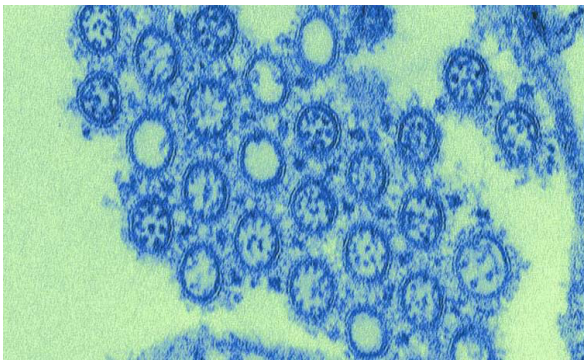
Imperial College London. Outbreak Analytics Workshop – Lusaka, Zambia
Wednesday 28th September 2022

Outbreak analysis practical: estimating the reproduction number

Dr Ilaria Dorigatti and Prof Steven Riley with contributions from
Dr Simon Cauchemez, Dr Anne Cori, Dr Patrick Walker Prof Christophe Fraser
Adapted by Rebecca Nash



2009 H1N1 influenza virus (source: CDC)

Background

Pandemics arise when a new virus capable of human-to-human transmission emerges that is sufficiently distinct from circulating viruses so that the level of population immunity is low or nil. New strains emerge by transfer from a zoonotic reservoir. Exactly how new viruses emerge is not clear and may differ for different pandemics. Possibilities include genetic re-assortment or recombination between human and avian viral strains, possibly via intermediary species (e.g. pigs, poultry or other animals), or by gradual accumulation of adaptive mutations.

The potential for pandemic viruses to cause significant mortality is illustrated not only by SARS-CoV-2, the causative agent of COVID-19, but also the 1918-20 H1N1 'Spanish flu' which is estimated to have killed at least 20,000,000 people worldwide. Note however that the much lower mortality caused by the 1957 H2N2 'Asian flu', the 1966 H3N1 'Hong Kong flu' or the 2009 H1N1 pandemics shows that devastation is not an inevitable result of a pandemic but depends on the biology of each new strain as well as the impact of interventions.

Preparedness is greatly helped by knowledge of the epidemiological determinants of virus spread, such as the basic reproduction number R_0 and the duration of infectiousness. The aim of this practical is to estimate some of these quantities.

Objectives

From a methodological perspective, the practical will introduce you to some of the methods and issues surrounding parameter estimation in epidemic models. More specifically, we are going to use *2 different methods* to estimate the reproduction number from epidemic data. In part 1 of the practical, we will use the tree reconstruction method called EpiEstim, and in part 2 we will estimate the reproduction number by fitting a compartmental SIR model to the epidemic curve.

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

We are going to consider data from a school outbreak of H1N1 pandemic influenza virus which occurred in Pennsylvania in 2009 (data from Cauchemez et al, PNAS, 2011; paper attached to practical).

Part 1

Using tree reconstruction methods to estimate the reproduction number

Fitting a transmission model to data imposes making a range of assumptions for example about the population size or the initial proportion of susceptibles in the population. Rather than making such assumptions, we are now going to use tree reconstruction methods described in the lecture to estimate the instantaneous reproduction number R_t in the school.

For this, we only need the epidemic curve and the distribution of the generation time. In this section we talk about serial interval as a synonymous with generation time. We will use the same incidence dataset as in section 1.

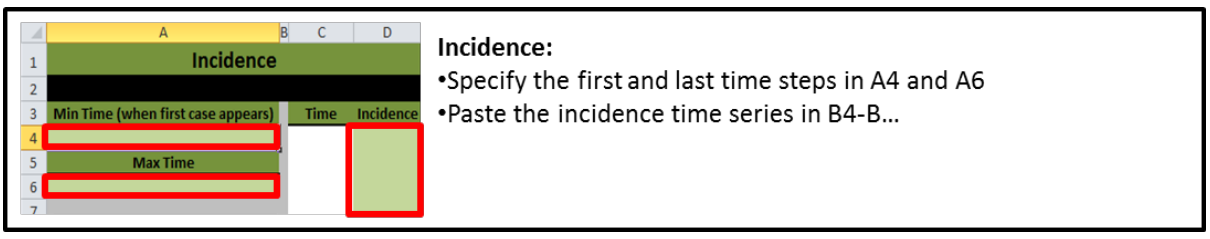
- Open the influenza.txt file to see the incidence data
- Open the Excel file EpiEstim.xls

There are several sheets in it: Readme, Data, Output1 serial interval, Output2 R estimates and Figures. Readme will provide you with information on how to use the document, which is summarized in the following.

Data is the only sheet you have to modify; only light coloured cells have to be modified.

1. **Fill in the Incidence section** as shown in Snapshot 1; this will be done only once as we will only look at one dataset. Take 1 and 32 as the Min and Max times to match the data.

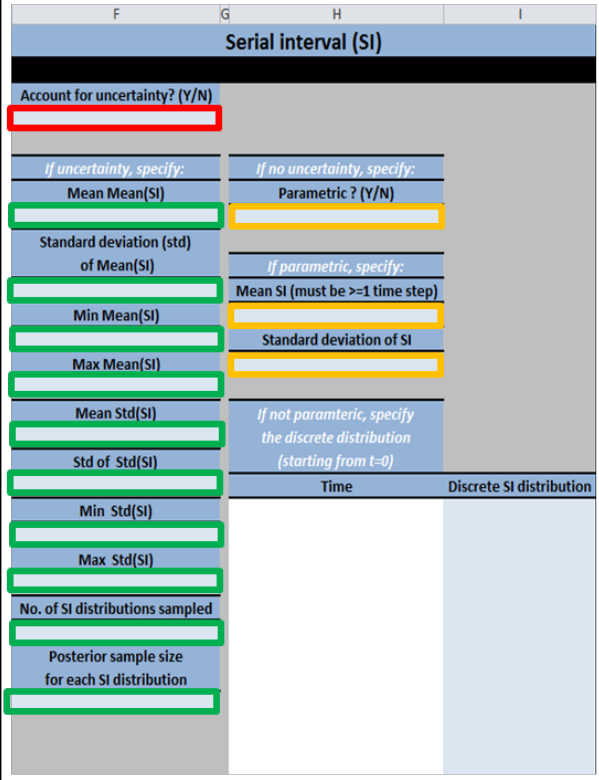
Snapshot 1:



2. **Specify your assumptions about the serial interval distributions** (see snapshot 2).

Snapshot 2:

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.



Serial interval (SI):

•Specify in F4 whether you want to account for uncertainty in the SI distribution (Y) or not (N)

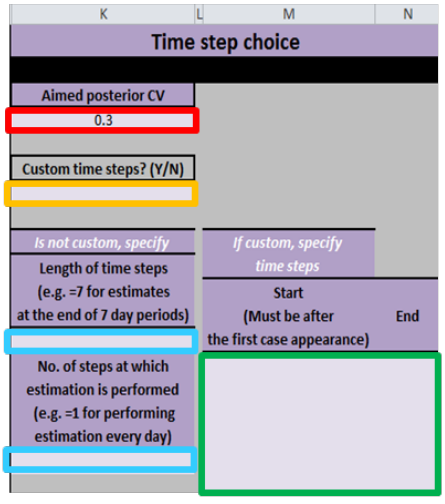
•First option (N): not accounting for uncertainty
Specify the SI distribution either in a parametric or non parametric way (H8)
→If parametric, provide mean and sd for the SI (H12 and H14)
→If non parametric, provide (in I20-I…) the whole distribution of the SI (time step given by data, starting at t=0)

•Second option (Y): accounting for uncertainty
Provide Mean, Sd, Min and Max
for the mean SI (F8, F11, F13, F15)
for the sd of the SI (F17, F19, F21, F23)
Provide in F25 the number of SI distributions to explore (50 by default, the bigger the longer the estimation)
Provide in F28 the number of R values to be drawn for each SI distribution explored (50 by default, the bigger the longer the estimation)

In a first analysis, we will use a FIXED parametric serial interval with mean 1.18 day and sd 0.96 day. Fill in the cells shown in red and yellow in Snapshot 2 accordingly.

3. **Specify the time windows** you want to use (see snapshot 3). Keep the posterior coefficient of variation to its default value of 0.3

Snapshot 3:



Time windows: Note that time windows can overlap!

•Specify in K4 the aimed posterior coefficient of variation (default 0.3, the smaller the more precise the R estimates but the longer the time windows need to be to get this precision)

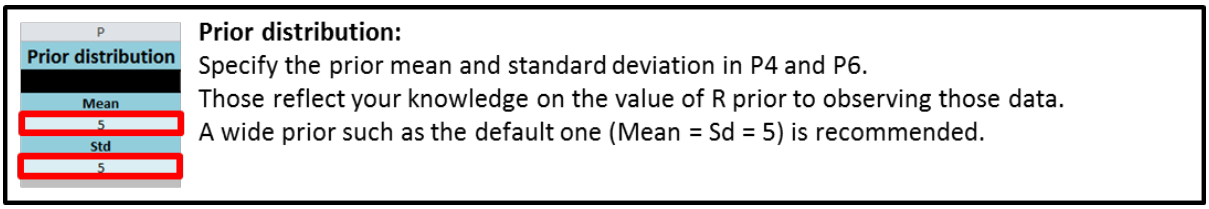
•Specify in K7 whether you want custom time windows (Y) or not (N).
•First option (N): custom time windows
Give start times in M14-M… and end times in N14-N…
•Second option (Y): non custom time windows
Give the windows length in K13 (eg 7 if daily data and weekly widows)
Give, in K18, the number of time windows at which estimation is performed (1 suggested to get estimates on sliding windows ending on each time step with data)

In a first analysis, we will explore the temporal variations of the reproduction number. For this, choose non-custom weekly sliding windows with a one-day lag between two successive windows (yellow and blue cells).

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

4. **Specify the prior mean and standard deviation** (see snapshot 4). Keep the default values of 5 and 5.

Snapshot 4:



5. **Enable Macros**

Click the File Menu and select Options from the left sidebar. In options, select Trust Center from the left sidebar and click Trust Center Settings button on the main window.

Now in Trust Center Settings dialog window, select Macro Settings from the left sidebar, choose Enable All Macros option and hit OK.

6. **Run the estimation!**

Snapshot 5:

7. **Results** are presented as tables in sheets “Output1 serial interval” and “Output2 R estimates” and as figures in sheet “Figures”.

What is your estimate of the initial R? How does it compare with the estimate of Cauchemez et al.? See highlight in Cauchemez et al, page 5.

- Mean 1.36, 95% credible interval 0.78-2.10

In Cauchemez et al: Mean 1.4, 95%CI 1.2-1.5

Our point estimate is similar, but we find more uncertainty. This could be due to us using only 1 week of data versus Cauchemez et al. using 2 weeks of data to derive the estimate.

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

When does your mean R estimate fall below the threshold 1?

The week from day 13 to day 19

What happens at the end of the epidemic (look at the third figure)? What is a possible explanation?

- The mean R estimate goes up at the end of the epidemic, because of the few late cases. Given the serial interval distribution it is very unlikely to have 4 days with no incidence and then 2 incident cases, unless R is very high. The explanation is that those cases were probably infected outside the school rather than in the school.

Impact of school holidays

The school closed during time interval day 18-day 24. Repeat the analysis using the following custom time windows (Change the yellow and green cells in Snapshot 3):

Day 1 to 17

Day 18 to 24

Day 25 to 31

What are your estimates for R before, during and after the holidays?

- Before: Mean 1.17 , 95% credible interval 0.95-1.41
- During: Mean 0.70 , 95% credible interval 0.46-1.00
- After: Mean 1.1 , 95% credible interval 0.36-2.25

What is the estimated reduction in the reproduction number during school closure compared to before school closure? How does that compare to the one obtained by Cauchemez et al (see highlight, page 4)?

Point estimate of reduction is (1.17-0.7)/1.17=40% here as opposed to 30% in Cauchemez et al – so they are roughly consistent. The slightly weaker effect in Cauchemez et al. could be due to us considering before versus during school closure and them considering (before+after) versus during closure.

Is the effect statistically significant? How does that compare to Cauchemez et al.?

- The two 95% credible intervals (0.95-1.41 before and 0.46-1.00 during school closure) are overlapping hence we don't find a significant effect. Cauchemez et al also found that the effect was not significant.

How do you think underreporting would affect R estimates? Would underreporting have an effect if the level of reporting was constant over time?

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

Changes in the reporting rates over time would bias the estimates.

However, the estimates would not be affected if underreporting is constant, regardless of the specific proportion of reported cases.

Part 2

Estimating the reproduction number by fitting a dynamical transmission model

In part 2 of the practical we are going to learn how to estimate the reproduction number using a compartmental model fitted to time series data.

Fitting a model and estimating parameters

For this analysis, we are going to make the following simplifying assumptions: 1) the outbreak in the school was closed (i.e. there were no importations and exportations of cases); 2) there was homogeneous mixing in the school; 3) there is no latent period and cases are infectious for the entire time they are infected; 4) we ignore issues of missing data and censoring in the data (see Cauchemez et al, PNAS, 2011 for an analysis of the data dealing with these problems).

Here, we will consider a simple SIR model:

$$N = S + I + R$$

Differential equations:

$$\frac{dS}{dt} = -\beta \cdot \frac{I}{N} \cdot S$$

$$\frac{dI}{dt} = \beta \cdot \frac{I}{N} \cdot S - \gamma \cdot I$$

$$\frac{dR}{dt} = \gamma \cdot I$$

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

Other helpful equations and how they can be rearranged:

$R_0 = \beta / \gamma$ OR $\beta = R_0 / D$ OR $R_0 = \beta * D$

$D = 1 / \gamma$ OR $\gamma = 1 / D$

Download the data and models

Access the odin app by following the link in the supporting files, or using the following link: https://shiny.dide.imperial.ac.uk/infectiousdiseasemodels-2021/flu/

Beneath the link to load the odin app in supporting files, click the link “Influenza data” which will download “influenza_data.csv”. Save this file in a convenient place. Then click the links “Pennsylvania, basic” and “Pennsylvania, with school closures” which should download “solution1.R” and “solution2.R” retrospectively and save these in the same folder as “influenza_data.csv”.

Run the model and link to the data

→ Open the app and move to the “Data” tab if not there automatically.

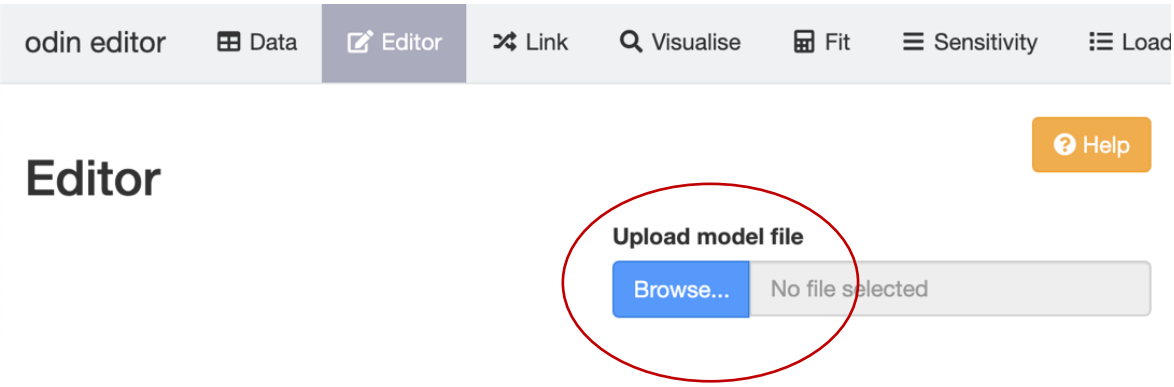
→ Click “Browse” and select “influenza.csv” in the location you saved it.

→ The data should appear as a graph and table.

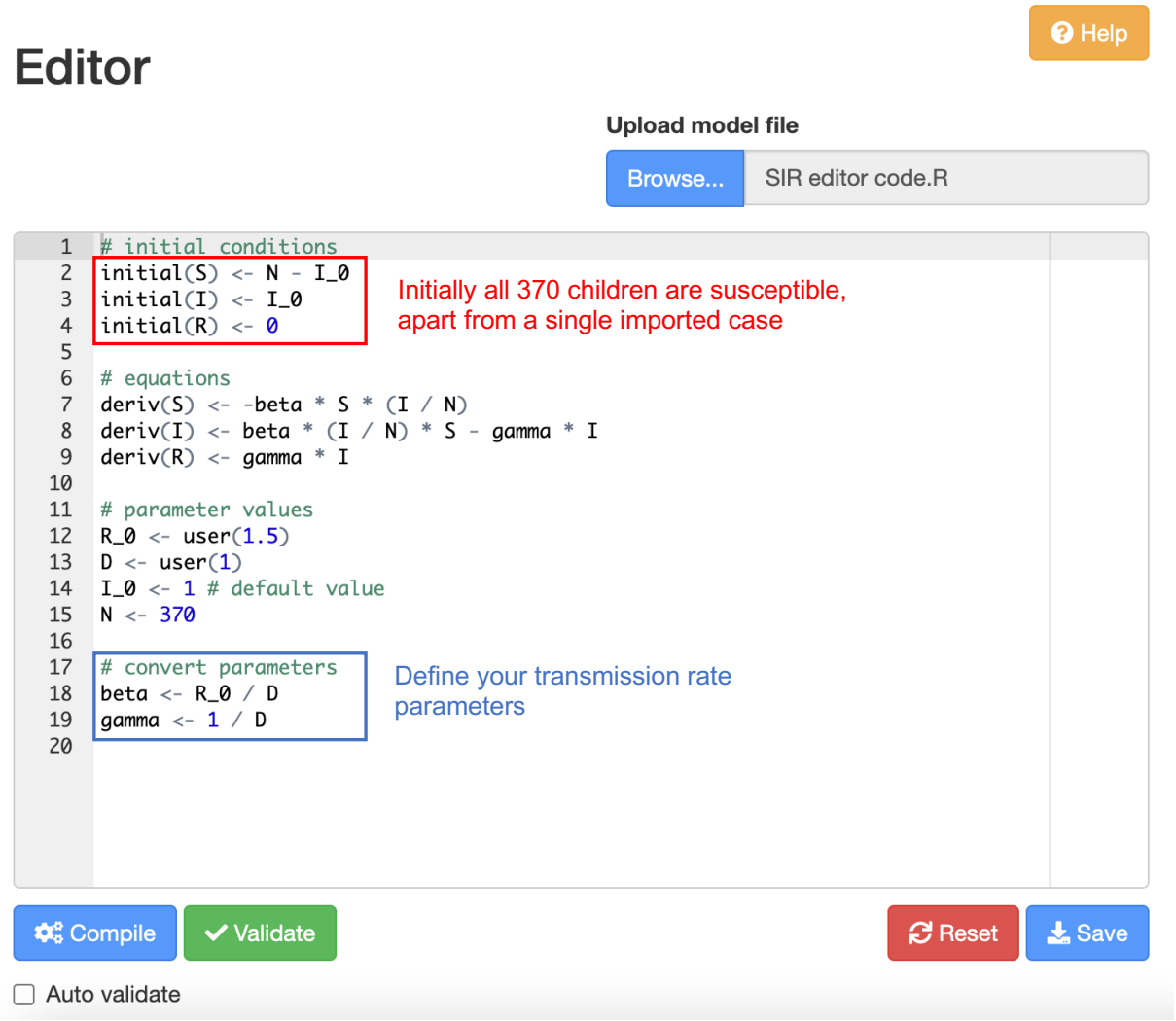
“Cases” gives the daily number of children of the school with symptoms onset.

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

→ Move to the “Editor” tab, ignore the code currently there and load “SIR editor code.R” via the “Browse” tab within the “Editor tab”.



→ The Editor tab should now look like the screen below:



Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

Look at how the SIR model is coded. What are the parameters D and I_0? (NB the model begins at Day 0 with the onset of the first case within the school").

D represents the mean duration of disease/infectiousness (here, remember we are assuming that cases are infectious for the entire duration that they're infected)

I_0 represents the initial infected case

Click "compile" to compile the model.

Now select the "Link" tab. Ensure there are tick marks to confirm the data has been successfully uploaded and model code has successfully been compiled.

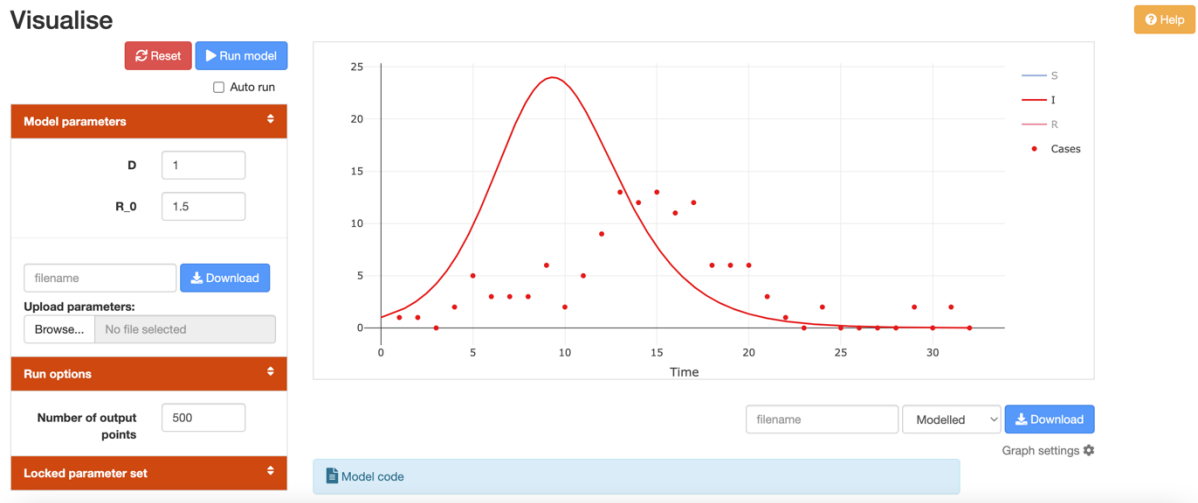
In the "Link" section we link the model to the data by specifying the variable in the model that is appropriate to fit to the data (i.e. "Cases"). Which variable is this?

I, as it describes the daily number of infected children who attend the school.

Now move to the "Visualise" tab. Guess a value for D and R_0 (all positive!) and click "Run model".

Manual fit

You should now see a chart of the dynamics of the variables of the model over time for your chosen set of parameter values, alongside the case data. How does your choice of parameter values match to the data? (NB it will help to visualise if you deselect all but the relevant). The picture below gives one example (of a poor fit!).



Vary the two parameters, how does each one affect the model output? Once you have a set of parameters which gives you a fit you're happy with (NB don't spend ages on this – if you haven't got a set you're happy with within, say, 5-10 minutes then just note down the set you think fits best of those you've explored so far).

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

(will vary from person-to-person but check that the values looks ok-ish) e.g.

R_0 = ~ 1.3

D = ~ 1.1 days

Move to the “sensitivity” tab and enter your best fitting set of parameters. Under “vary parameter” select each parameter in turn and select “range”, specify a reasonably wide range of values for each parameter and 10 runs. Click “Run model”. Can you describe the effect each parameter has on your output? Understanding this may help you to find a set of parameters you are happy with…

Varying D changes the width of the peak. As D increases the generation time increases, leading to slower spread for the same R0.

For R0, after passing the threshold value of 1 epidemics begin to occur. Epidemic peaks higher but also peaks earlier as R0 increases. Not easy to see on the plot but the final size of the epidemic (total # infected before epidemic dies out) also increases with R0 but the epidemic finishes much more rapidly due to a faster depletion of susceptibles.

Automated fit

Now we will assess the extent to which we are as skilled as the computer in finding a good fit to the data. In the “Fit” tab enter the set of parameter values you think gives you the best fit. This will replot the data, in the tab alongside a number called “sum of squares” – this is a weighted sum of the square of the differences between each data point and the variable you are fitting (i.e. if they were exactly the same the sum of the squares would be zero indicating a perfect fit).

If there are more than one set of parameters you think give an equally good fit (or you had a different set before and after you visited the ‘sensitivity’ tab) compare the sum of squares and set the values to those that give you the smallest number (as this is the one closest to going through all of the data points). Now make sure R_0 and D are all ticked (to include them in the fitting process) and click ‘Fit model’. Note down the fitted sum of squares.

Sum of squares =

Auto fit parameters:

R_0 = 1.27

D = 1.03 days

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

The relationship between the epidemic growth rate r, the basic reproduction number R_0 and the generation time T_g.

The generation time T_g is the time it takes between infection of one individual and subsequent infection of others by that individual.

If the epidemic growth rate r and the generation time T_g are known, it is possible to derive R_0 from the following equation:

$$R_0 = \frac{1}{\int_0^\infty \exp(-r.t)\, g(t)\, dt} \qquad (2)$$

where $g(t)$ is the generation time distribution (i.e., the probability that the generation time is equal to t).

For an SIR model, if the growth rate (r) and D are known, the equation for R_0 can be simplified as follows:

$$R_0 = 1 + rD$$

If R_0 and D are known, this can be rearranged to estimate r:

$$r = (R_0 - 1)/D$$

And if r and R_0 are known, the generation time (T_g) can be estimated:

$$T_g = (R_0 - 1)/r$$

Based on this knowledge, what is the generation time, T_g, of the best-fitting model?

r = (1.27 – 1) / 1.03 = 0.26

Tg = (1.27 – 1) / 0.26 = **1.04**

Why does the generation time in the school appear to be shorter than the generation time in the household? See highlight in Cauchemez et al, page 2.

Because transmission in the school is truncated compared to what it is in the household (mixing of infected pupils reduced by e.g., pupils staying away from school when symptomatic).

How do your estimates of R and the generation time compare with those of Cauchemez et al? See highlight in Cauchemez et al, page 5. How does it compare with your estimate from EpiEstim?

R=1.3 here compared to 1.4 in Cauchemez et al.

Tg ~ 1 here compared to 1.5 in Cauchemez et al.

The initial EpiEstim R estimate was 1.36.

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

Impact of school holidays

The school closed during time interval day 18-day 24. We are going to try to estimate the impact of school closure.

Head to the "Editor" tab, hit "Browse" and now select "school holidays.R" you downloaded earlier. This should then display code that has been modified so that the reproduction number is different when the school is open (R_0) and when the school is closed (R_0_closure).

Hit compile! Then link the model and data in the "link" tab

The "Visualise" and "Fit" tab should reflect the new model with the additional parameter. If not retrace your steps from the "**Run the model and link to the data"** section previously.

Play around with this model to your hearts delight then redo the auto-fitting process, making sure you now tick three parameters (D, R_0, and R_0_closure). Repeat this a couple of time with different initial guesses to check the computer has identified a good solution (if different values, again use the one with lowest sum of squares):

Sum of squares = 102.3425

R_0 = 1.29

R_0_closure = 0.64

D = 1.27 days

What is the estimated reduction in the reproduction number during school closure? How does that compare to the one obtained by Cauchemez et al (see highlight, page 4) and the estimates from EpiEstim?

Point estimate of reduction is (1.29-0.64)/1.29=50% here opposed to 30% in Cauchemez et al and 40% when using EpiEstim

Is the effect of school closure statistically significant?

There is an improvement in the sum of squares when school closure is accounted for. However, although odin is a nice tool to quickly explore model dynamics and fit models to data, it only provides point estimates. It does not provide a confidence interval for parameters (possible – but difficult to do). So here for example, we cannot determine whether the effect was significant. Cauchemez et al found that the effect was not significant (70%; 95% CI: 38%, 122%).

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.

What are the aspects of the outbreak investigation that you haven't accounted for and that may lead to bias estimates?

- Many individuals were only interviewed after the peak. Observed decay after peak may therefore be sharper than what happened in reality.
- Possible superspreading event early in the epidemic.
- In practice, it didn't appear to be homogeneous mixing. There was clustering of classes in classes. 4th graders were more affected for example.

Summary and discussion

After trying out these 2 methods of estimating R (EpiEstim vs model fitting), can you think of pros and cons of the different methods? What are the key sensitivities?

EpiEstim vs model fitting:

- pros: less assumptions on mechanisms, often more robust

- cons: can't distinguish between effect of reduction of susceptible population and reduction of transmission due to interventions

To estimate R0 from an epidemic curve, we need to make an assumption on the distribution of the generation time. Estimates are sensitive to the assumption made – they depend on the mean, but also on the shape and variance of the distribution of Tg.

Copyright © Imperial College London 2020. All rights reserved. Not for reproduction or distribution.