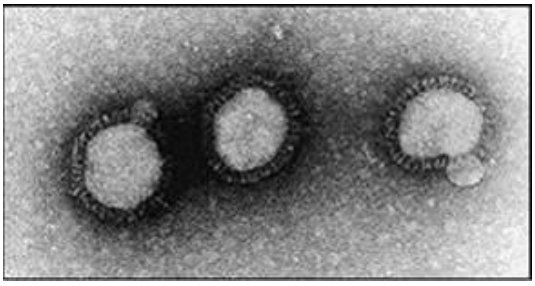
Imperial College London
9 November 2018

Practical: Estimating key epidemiological determinants of pandemic influenza

Ada Yan and Helen Coupland

Original material by Dr Ilaria Dorigatti, Dr Simon Cauchemez, Dr Anne Cori, Prof Steven Riley & Prof Christophe Fraser



The influenza virus

Background: Pandemic influenza

Pandemics of influenza arise when a new strain capable of human-to-human transmission emerges that is sufficiently distinct from circulating strains so that the level of population immunity is low or nil. New strains emerge by transfer from aquatic birds, which are the natural reservoir of influenza. Exactly how new strains 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 (pigs, poultry, etc…), or by gradual accumulation of adaptive mutations.

The potential for pandemic influenza to cause significant mortality was illustrated by the 1918-20 H1N1 'Spanish flu' which is estimated to have killed at least 20,000,000 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.

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

Objective: Estimating transmission parameters

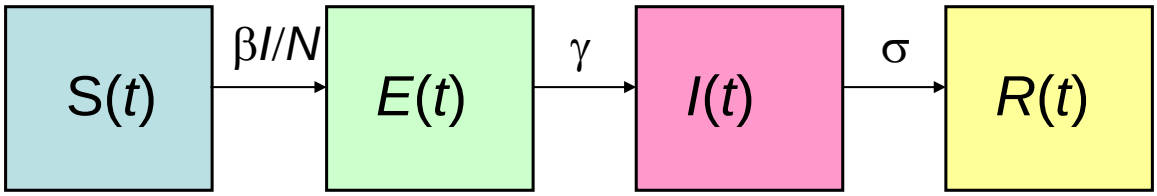
From a methodological perspective, the practical will introduce you to the methods and issues surrounding parameter estimation in epidemic models. More specifically, we are going to use 2 different methods to estimate the reproduction number from epidemic data. In the first part of the practical, we had focused on the analysis of exponential growth rates. In this second part we will see how to obtain estimates of the reproduction number by fitting a compartmental model to the epidemic curve.

We are going to consider a dataset which describes an A(H1N1)pdm09 pandemic influenza school outbreak in Pennsylvania in 2009 (Cauchemez et al, PNAS, 2011; paper attached to the practical).

Transmission of A(H1N1)pdm09 pandemic influenza in a school in Pennsylvania

We are going to fit a simple transmission model to data collected during a school outbreak of A(H1N1)pdm09 influenza. For this analysis, we are going to make the following simplifying assumptions: 1) the outbreak in the school was closed; 2) there was homogeneous mixing in the school; 3) 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 SEIR model, written as a set of ordinary differential equations (ODEs):



$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dE}{dt} = +\frac{\beta SI}{N} - \gamma E$$
$$\frac{dI}{dt} = \gamma E - \sigma I$$
$$\frac{dR}{dt} = \sigma I$$
$$N = S + E + I + R$$

β, γ and σ are rates of processes in the model: $\frac{\beta I}{N}$ is the force of infection, γ is the rate at which individuals become infectious, and σ is the recovery rate. It may be useful to express these quantities in terms of more familiar quantities:

$$R_0 = \frac{\beta}{\sigma}$$
$$L = \frac{1}{\gamma}$$
$$D = \frac{1}{\sigma}$$

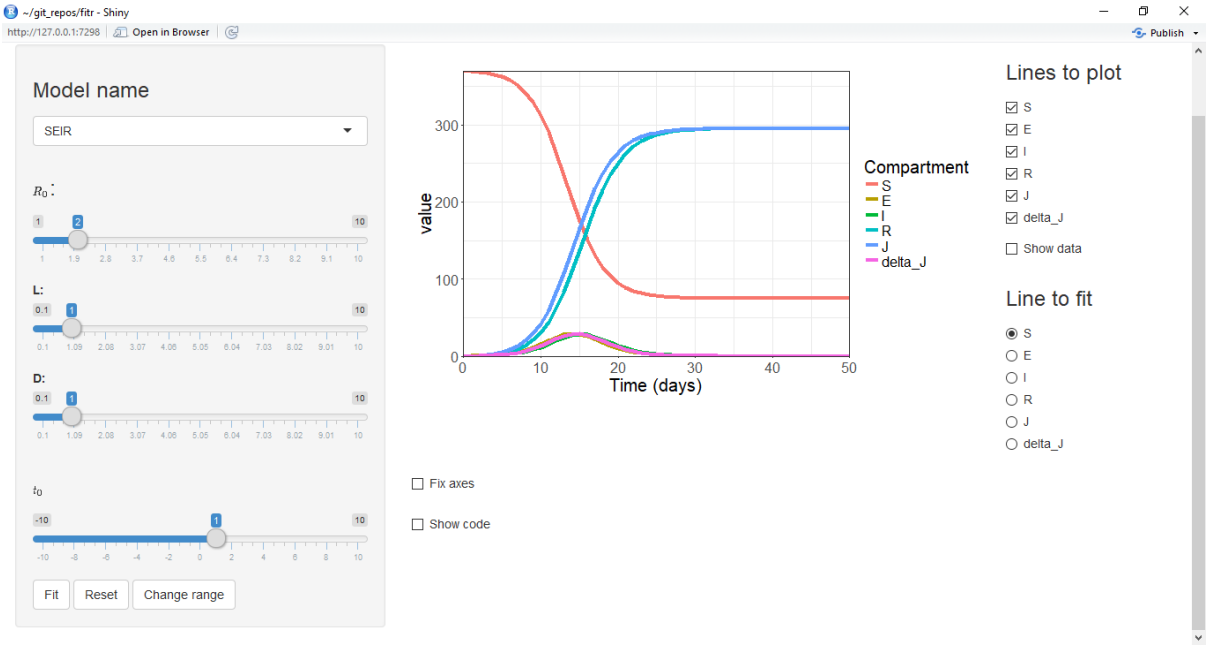
Where R_0 is the basic reproduction number, L is the mean latent period and D is the mean duration of infectiousness. The mean generation time for this model is $T_G = L + D$.

An ODE model requires a set of initial conditions. In our case, we assume that one individual is exposed at time t_0, and all other individuals are susceptible at this time. Hence, the initial conditions are

$$S(t_0) = N - 1$$
$$E(t_0) = 1$$
$$I(t_0) = 0$$
$$R(t_0) = 0$$

Understand model behaviour

Go to https://shiny.dide.imperial.ac.uk/fitr/ in your web browser.



The model solution automatically updates as you change the current parameter values. Select S, E, I, R from "Lines to plot" and change the parameter values one by one.

What happens to the epidemic as each of the parameters is changed?

It may be helpful to select "Fix axes" and/or remove certain lines from the plot.

As R_0 increases, the total number of infected individuals infected by the end of the epidemic increases – R reaches a higher value, S reaches a lower value, and the areas under the E and I curves increases.

As L and D increase, the epidemic progresses more slowly. Increasing L increases the time that individuals spend in the incubation period compared to the infectious period, so the area under the E curve increases relative to the area under the I curve.

As t_0 increases, the first case is introduced later, so the whole epidemic shifts to the right.

Plot the data

We have daily incidence data for the outbreak. Check the box "Show data".

Which compartment out of S, E, I, R should we fit the incidence data to, if any? Why?

None – the daily incidence is the number of individuals entering the I class each day, which is not currently represented by any of the equations.

To fit the model to incidence data, we need to calculate the daily incidence. We modify the model to include a differential equation for cumulative incidence J. The rate of change of the cumulative incidence is the rate as which individuals enter the infectious compartment.

$$\frac{dJ}{dt} = \gamma E$$

The daily incidence for day t is then the cumulative incidence on day t, subtracting the cumulative incidence on day $t-1$.

$$\Delta J(t) = J(t) - J(t-1)$$

Adding these compartments to the model doesn't change the equations for S, E, I and R, so the epidemic dynamics stay the same.

Check "Show code". Look at the code for the ODEs, and make sure that you understand each line.

Manual fit

You are going to try to find the set of parameters that gives the best fit to the data.

Select "delta_J" only from "Lines to plot". Using parameter sliders, manually modify parameter values to find the set of parameters that gives the best fit to the data. Note results below [NB: this manual fit exercise is essentially for you to gain a sense of how the fit is influenced by the values of the different parameters – so, don't worry if you are not completely satisfied with the fit you get and think you could do better].

R_0	= 1.25	
L	= 0.47	days
D	= 0.4	days
t_0	= 0.7	days

Maximum likelihood parameter estimation

We will find the parameter set that maximises the likelihood of the data given the parameters. We assume that each point in the data is Poisson distributed, with mean equal to the value predicted for the compartment indicated by "Line to fit". We assume that each observation is independent. The overall likelihood is thus the product of the likelihoods of each point given the model parameters.

The log likelihood is the natural logarithm of the likelihood. The likelihood can take very small values for some parameter sets, so we maximise the log likelihood instead. This is a technical point – maximising the log likelihood also maximises the likelihood.

Select "delta_J" as the line to fit, and vary the parameters using the sliders again.

What happens to the log likelihood as the incidence curve moves away from the data?

The log likelihood decreases, indicating that the fit is worse.

Move the slider to t_0 = 3 days. Why is the log likelihood -Inf at this point? (When the log likelihood is -Inf, the likelihood is zero.)

The Poisson distribution does not allow for false positives. If the first case is introduced at t = 3 days, the probability of the observed incidence being greater than zero before that time is zero, so the overall likelihood is zero.

The optim function in R finds the parameter values that maximise (or minimise) a function. The Fit button runs the optim function behind the scenes. The optim function requires an initial guess for the parameter values. It also requires a range of sensible parameter values within which to search.

Click "Change range" to specify a range of sensible parameter values within which to search. Then use the sliders to make your best guess for the parameter values.

→ *Click Fit to start the automated fitting procedure.*

Note the estimates you get below:

Log likelihood	= -62.5	
R_0	= 1.25	
L	= 0.47	days
D	= 0.4	days
t_0	= 0.7	days
T_g	= L + D = 0.87	days

[In general, it is safer to repeat the procedure several times with different initial guesses. If the fit improves (higher log likelihood), record the new parameters – these are a better fit of the model to

the data. There is no absolute way of making sure you have found the best possible fit – numerical optimisation requires a combination of good programming plus trial and error. Also take care in defining the range for the optimisation.]

How do your estimates of R_0 and the mean generation time compare with those of Cauchemez et al? See highlight in Cauchemez et al, page 5.

R_0 = 1.3 here compared to 1.4 in Cauchemez et al.

T_G = 0.9 here compared to 1.5 in Cauchemez et al.

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

Because students may stay home from school when symptomatic, reducing the potential period for transmission.

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.

Check to box "school closure". The code has been modified so that the basic reproduction number is different when the school is open (R_0) and when the school is closed ($R_{0closure}$). Look at the code to see how this has been done.

Fit the new model to the data:

Log likelihood = -57.3

R_0 = 1.32

$R_{0closure}$ = 0.86

L = 0.47 days

D = 0.69 days

t_0 = 0.5 days

T_g = 1.16 days

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

(1.32-0.86)/1.32 = 35%. Consistent with the 30% by Cauchemez et al.

Is there evidence that the basic reproduction number changed during school closure?

$$-2\Delta LL = 2 \times (-57.3 - -62.5) = 9.8$$

df = 1 because the model with school closure has one more parameter.

p-value = pchisq(9.8, 1, lower.tail = FALSE) = 0.002

For a significance level of 0.05, we reject the null hypothesis that the model with school closure does not provide a better fit. This result supports the hypothesis that R_0 changed during school closure. However, the maximum likelihood fitting procedure we have used does not provide confidence intervals, so we don't know how different R_0 was. Cauchemez et al. found that the 95% confidence of the ratio of $R_{0closure}$ to R_0 spanned 1 (95% CI: 38%, 122%), so the difference was not statistically significant.

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

- We assumed this was a closed outbreak but there might have been interactions e.g. with family members.
- 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.