MATH35600 Assessed Practical 2018

You should work in groups of 3 for this assignment, handing in one report at the end. It is worth 20% of the marks for this unit.

The data loaded with the command

```
h1n1 <- read.table("https://people.maths.bris.ac.uk/~sw15190/TOI/sh-H1N1.dat")
```

give the number of laboratory confirmed `cases` of H1N1 influenza in the Southern Hemisphere each `week` starting in the 17th week of 2009. The data cover the period of the southern hemisphere winter and are from the WHO. It may be reasonable to assume that the expected number of lab confirmed cases is proportional to the actual number of cases each week, but of course most cases will not be lab tested. One possible model for these data is that they result from an S-I-R (susceptible-infected-recovered) model, of the form

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\alpha SI, \qquad S(0) = 1 \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \alpha SI - \beta I, \quad I(0) = I_0 \end{aligned}$$

$S(t)$ is the number of individuals *susceptible* to H1N1, scaled by the initial number susceptible at time 0. $I(t)$ is the number of individuals with the disease and *infective* at time t (also scaled by the initial number of susceptibles). α is the transmission rate parameter, and β the recovery rate parameter. I_0, the initial number of infectives, is also an unknown parameter. Time 0 can be taken as the start of the observed data. The equation for *Recovereds* is not shown as it is irrelevant here. Exactly how the observed confirmed cases, y, are related to I is not quite clear. One possibility is that

$$y_t \sim \mathrm{Poi}\{\phi I(t)\},$$

but another is that

$$y_t \sim N\left(\phi I(t), \gamma\sqrt{\phi I(t)}\right),$$

where ϕ and γ are unknown parameters. It could be that neither distributional model is right (although you should assume that $E(y_t) = \phi I(t)$).

It is of particular interest to know whether there is evidence for seasonal change in the parameter α in the southern hemisphere (i.e. whether α changes in time, possibly increasing in the colder months).

1. The first task is to write code to find approximate solutions to the S-I-R model. For the current task it is best to do this by a simple weekly discretization of time. Normally the simplest approximate solution method for a differential equation model is Euler's method. If we write the model as

   $$\frac{\mathrm{d}N}{\mathrm{d}t} = a_t$$

   we pretend that a_t is constant over some timestep Δt, and hence obtain

   $$N_{t+\Delta t} = N_t + a_t \Delta t,$$

   which is iterated to get an approximation to the solution of the original model.

   In the current context Euler's method has the disadvantage that it can produce negative solutions, which are here impossible. Hence it is better to approximate the dynamics over Δt by an approximating exponential growth or decline model. That is we write the model in the form:

   $$\frac{\mathrm{d}N}{\mathrm{d}t} = r_t N(t)$$

and then treat r_t as fixed over $[t, t + \Delta t)$. This results in the approximation

$$N_{t+\Delta t} = N_t \exp(r_t \Delta t)$$

which is iterated to find the approximate solution to the original model. For example, $\mathrm{d}N/\mathrm{d}t = rN(1 - N/K)$ can be approximated by $N_{t+1} = N_t \exp\{r(1 - N_t/K)\}$.

Apply this approach to each of the two equations making up the proposed SIR model, to arrive at a pair of discrete time equations, with a one week time step, approximating the original model

Notice that by treating the equations separately, your approximate model does not force the number of individuals leaving the susceptible class to match the number entering the infective class.

Work out how the following discretisation is obtained

$$\begin{aligned} S_{t+1} &= S_t e^{-\alpha I_t}, & S(0) &= 1 \\ I_{t+1} &= S_t(1 - e^{-\alpha I_t}) + I_t e^{-\beta}, & I(0) &= I_0 \end{aligned}$$

Why is preferable to your original discretisation? Write an R function which will take model parameters as input, and return a vector of weekly predicted I values as output.

2. Decide on a reasonable distributional assumption for the observed case data, write an R function to evaluate the log likelihood of the model parameters, and hence find the MLEs of the SIR model parameters. Make sure that you check the distributional assumption, and modify it if necessary.

3. Investigate whether the data provide any evidence that the infection rate, α, varies in time. To do this create models in which the fixed parameter α is replaced by a linear or quadratic function of time (with parameters to be estimated), and compare these models to the model with α fixed in time.

What to hand in: You should write, as a group, a concise report describing the model estimation process, and critically discussing the model fits, including what they say about the adequacy of the model. Report your biological conclusions. The report should include appropriate plots. It should describe what you have done in sufficient detail that a competent statistician can follow both what was done, and why, and can thereby form a judgement as to the reasonableness of the analysis. The report should focus on using statistical inference theory to investigate these data, not on deriving that theory.

The written report should be no more than 5 sides of A4 (normal margins $\geq$ 10pt font). The report should be accompanied by an appendix containing well structured, clearly commented R code for performing the analysis.

One report (as a pdf file) and one appendix (plain text file is best) per group should be emailed to `simon.wood@bristol.ac.uk` with the subject `MA35600 H1N1` followed by your surnames, by **12 noon, on Friday 20th April 2018**. The deadline reflects the difficulties occasioned by the current industrial action, not the length of time that this should take!

Mark scheme guidance

First class marks will be awarded for work that could be passed on to the scientists who gathered the data essentially without modification. That is to say the statistics is appropriate and clearly explained, the conclusions appropriately drawn and any limitations are discussed fairly.

Upper second class marks will be awarded for work that could be passed on to the scientists who gathered the data, after a round of revision correcting some errors of presentation, interpretation or statistics that are relatively minor.

Lower second class marks will be awarded to work that has some more substantial flaws of presentation, interpretation or statistical reasoning which would require some more work to correct.

Third class marks will be awarded for work that contains some indication of substantive understanding and engagement, but contains more serious errors and misunderstandings.