Applied Statistical Modelling (STAT 40510)
Main Project
Task 2: Model Fitting

Dr Lennon Ó Náraigh

TBC

Format of the Project

The main project in Epidemiological Modelling in STAT 40510 will be made up of several tasks.

- Follow the online lectures independently, attend weekly office hours in Weeks 5-7.
- Over the same time period, complete (in a group) **Tasks 1 and 2** to test your knowledge of what you have learned.
- Again over the same time period, you will be assigned your most challenging task, **Task 3**. You should begin to do background reading to understand what is required here.
- In Week 8, you should present your work to date, the presentation should consist of:
  - The theoretical concepts you have learned in Tasks 1–2;
  - How you will apply these in Task 3.
- The final report (due towards the end of the trimester) will be based entirely on Task 3.

In the class notes, we used nonlinear least squares to fit an SIR model to some data. This was done in a rather crude way, we simply wrote down a good cost function (found by trial and error):

$$J(\beta, \gamma, t_{offset}) = \sum_{i=0}^{N} \left[\log(I_{model}(t_i) + \epsilon) - \log(I_{data}(t_i) + \epsilon)\right]^2,$$

where $I(t_i)$ is the number of infectious people on day t_i, with the subscripts denoting the value of $I(t_i)$ in the model and the data, respectively. We then computed the optimal model parameters as:

$$(\beta^*, \gamma^*, t^*_{offset}) = \arg \min_{\substack{\beta \geq 0, \\ \gamma \geq 0, \\ t_{offset} \leq 0}} J(\beta, \gamma, t_{offset}).$$

A better way to fit model parameters to the data is to use the maximum likelihood function. The aim of this project is to learn about this approach, by applying an SEIR model to a dataset for influenza.

1 The dataset

We look at the outbreak of Influenza in Cumberland (USA) in 1918. The population is $N = 5234$. The number of cases counted per week is recorded and given in Table 1. We fit an SEIR model to the data as follows. We have:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta}{N} SI, \tag{1a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\beta}{N} SI - aE, \tag{1b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = aE - \gamma I, \tag{1c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I. \tag{1d}$$

In this context, 'R' means immune or recovered. We take $t \geq 0$, and we measure time in days. For the initial conditions we take (as given in the data):

$$S(0) = 0.7N, \qquad E(0) = 0, \qquad I(0) = I_0, \qquad R(0) = 0.3N - I_0,$$

The number of people becoming infected at time t is equal to the number of previously-exposed people entering the 'I' compartment at that time, hence aE. The number of new recorded cases is C, there will be $\mathrm{d}C/\mathrm{d}t$ new infectious cases per unit time, this is proportional to the number of people leaving the exposed compartment, hence:

$$\frac{\mathrm{d}C}{\mathrm{d}t} = faE.$$

Here, f is a number between 0 and 1: $f = 1$ corresponds to all infected people being recorded as such by the authorities, which would happen with universal testing or diagonsis. In this way, the number of new infections recorded in a time interval from τ_1

Week 1	6
Week 2	9
Week 3	20
Week 4	40
Week 5	77
Week 6	138
Week 7	219
Week 8	287
Week 9	294
Week 10	235
Week 11	153
Week 12	87
Week 13	45
Week 14	23
Week 15	11
Week 16	5
Week 17	2
Week 18	1
Week 19	0

Table 1: Number of new cases recorded in each week of the outbreak. Week 1 ends 7th September 1918.

to τ_2 is:

$$C(\tau_2) - C(\tau_1) = fa \int_{\tau_1}^{\tau_2} E(t)\,\mathrm{d}t.$$

We interpolate the results of solving the ODE onto a discrete time points $t_0, t_1, \cdots$, where $t_i = i$ Days. The number of newly recorded infections in the first week (7 days) is thus:

$$C(t_7) = fa \int_0^{7\text{ Days}} E(t)\,\mathrm{d}t.$$

We therefore introduce notation – and a formula – for the number number of recorded infections in the j^{th} week:

$$C_j \approx fa \sum_{i=7(j-1)+1}^{7j} \left(\frac{\mathrm{d}C}{\mathrm{d}t}\right)_{t_i} \Delta t, \qquad \Delta t = 1 \text{ Day}, \qquad j \text{ in weeks} \tag{2}$$

The aim now is to fit the data in Table 1 to the model prediction for the weekly new case count C_j in Equation (2).

2 Least Squares Again

We fit the model (Equation (2)) to the model by introducing the cost function:

$$J(\boldsymbol{x}) = \sum_{j,\ \text{weeks}} \left(C_j^{\text{model}} - C_j^{\text{data}}\right)^2, \tag{3}$$

where $\boldsymbol{x} = (a, \beta, \gamma, I_0, f)$. Sample results are shown in Figure 1.

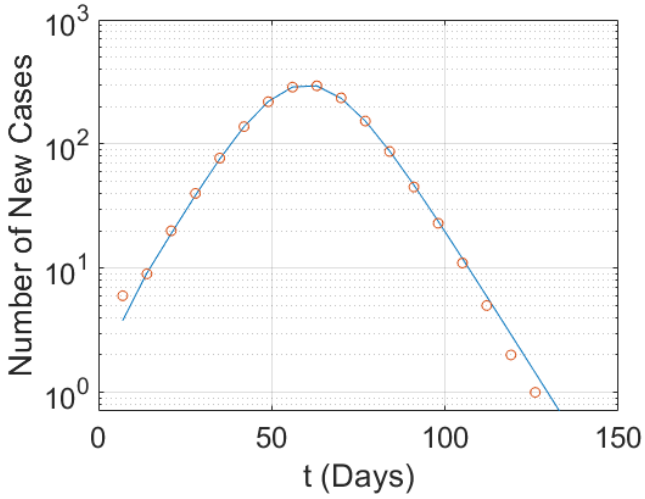

Figure 1:

3 Maximum Likelihood Function

Rather than least squares (where the choice of cost function is a bit arbitrary), it is better to fit the model to the data using the Maximum Likelihood function. We again look at the case count C_j in Equation (2). We assume that the observed case count C_j^{data} at time t_j (in weeks) satisfies:

$$C_j^{\text{data}} = C_j^{\text{model}} + \epsilon_i,$$

where ϵ_i is an error term.

Maximum Likelihood estimation requires that we make some assumption for the distribution of the error terms. Here, we assume that the error terms are independently normallly distributed. The maximum likelihood function is the 'probability of the observations given the model', hence:

$$L(\boldsymbol{x}) = \prod_{j,\ \text{Weeks}} p(C_j^{\text{data}}; \mu_j, \sigma(\boldsymbol{x})),$$

where:

- p is the normal distributioin;
- μ_j is the model value, $\mu = C_j^{\text{model}}$;
- The variance is:

$$\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}[C_j^{\text{data}} - C_j^{\text{model}}]^2}$$

  where N is the number of weeks of observations.

We then minimize the negative of the log-likelihood function:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \left[-\log L(\boldsymbol{x})\right],$$

which corresponds to maximizing the log-likelihood function.

Remark

With ϵ_i normally distributed, ML is equivalent to the least-squares minimization (3). However, ML has a better theoretical basis, the choice of the cost function is less arbitrary. Another benefit is that the modeller can try different options for the error terms. As well as a model where the ϵ_i's are normally distributed, other models can be tried as well, for instance a Poisson process.

In the Poisson process, what is usually observed is the cumulative case count, in our case:

$$CM_j = \sum_{i=1}^{j} C_i, \qquad j \text{ in weeks.}$$

Then, the observed cumulative case count is described as a Poisson process, the j^{th} datapoint CM_j^{data} is modelled as a Poisson process with Poisson parameter CM_j^{model}. The likelihood function to minimize is thus:

$$L(\boldsymbol{x}) = \prod_{j,\ \text{Weeks}} p(C_j^{\text{data}}; \lambda_j),$$

where $\lambda_j = C_j^{\text{model}}$ and $p(n, \lambda) = (\lambda^n / n!)\mathrm{e}^{-\lambda}$.

Finally, using ML estimation, the variances can be estimated from the Fisher Information matrix F, where

$$H_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j} \left[\log L(\boldsymbol{x})\right], \qquad F = H^{-1}.$$

4 The task

Using Least Squares and ML estimation, fit an SEIR model to the data in Table 1. Construct confidence intervals using both bootstrapping and the Fisher Information matrix. Comment on the results.