SIR Model and Nonlinear Least Squares

Brian Hunt
University of Maryland
AMSC/MATH 420, Spring 2015

More Sophisticated Models

- Let’s re-examine the assumptions behind our first models and discuss how to make them more realistic.
- We assumed a fixed population size N that was isolated from other sources of the hypothetical illness we modeled.
- We assumed that a single number p represents the probability of an infectious person infecting a susceptible person on each day, for each such pair of people.
- A more realistic model would allow p to depend on a number of factors.

Modeling the Infection Probability p

- In real life, the infection probability p depends on the pair of people. However, introducing an independent probability p_{mn} for each pair of people m and n results in way too many parameters.
- Also, p depends on time; for example, day of week.
- Perhaps most importantly, p depends on how long the infectious person has had the illness. Typically it peaks a certain amount of time after infection, then decreases to 0.
- To keep the number of parameters manageable, we need to have a model for how p depends on these factors.

Compartmental Models

- Many models divide the population into a relatively small number of categories (“compartments”) and keep track of the number of people in each compartment.
- Our first deterministic models had two compartments: “susceptible” and “infectuous”. We’ll call the continuous time model (3) the SI model.
- A widely studied model is the SIR model, which introduces a third compartment: “recovered” or “removed”. People in this category are no longer infectious.
- Other possible compartments can take into account more stages in the progression of the illness, different behavior patterns, different biological characteristics, etc.

Fitting to Data, Revisited

- In our earlier discussion, we assumed that the number of infectious people at a given time could be measured. But how would we ever know this number?
- The number of infectious people is often inferred from data on new diagnoses of the illness. However:
  - Not all people who get the illness see a doctor.
  - Diagnosis may come well after a person becomes infectious.
  - Data is not always reported (e.g., to CDC) promptly or reliably.
- A common problem in modeling is to relate the quantities of interest to the available data.

SI Model

- The SI model we discussed before is often written

  $$dS/dt = -pS\mathcal{I}$$
  $$d\mathcal{I}/dt = pS\mathcal{I}$$

  where S is the "susceptible" population – those at risk to become infected at a given time – and $\mathcal{I}$ is the infectious population. For this model the sum $S + \mathcal{I}$ remains constant over time; we called the sum N and substituted $S = N - \mathcal{I}$ in the second equation.
- The resulting solution was

  $$\mathcal{I}(t) = \frac{N\mathcal{I}(0)}{\mathcal{I}(0) + [N - \mathcal{I}(0)]e^{-pNt}}$$

SIR Model

- The SIR model (Kermack & McKendrick, 1927) is
$$
\begin{aligned}
dS/dt &= -pS\mathcal{I} \\
d\mathcal{I}/dt &= pS\mathcal{I} - r\mathcal{I} \\
dR/dt &= r\mathcal{I}
\end{aligned}
$$
  where R (for “recovered” or “removed”) is the number of people who were infected but are no longer infectuous. In this case, $\mathcal{I} + R$ is the cumulative number of people infected.
- One can add a term to the first equation representing new arrivals to the susceptible population.
- There is no formula for the solutions.

Properties of Solution Families

- Each model's family of solutions has some properties that are useful for fitting parameters to data.
- A time-shifted solution is also a solution: If $\mathcal{I}(t)$ is a solution, then $\mathcal{I}(t + c)$ is also a solution (with a different initial condition). This is because the model is "autonomous" – no explicit t dependence.
- A rescaled solution is also a solution: If $\mathcal{I}(t)$ is a solution, then $a\mathcal{I}(bt)$ is a solution of the same model with different parameters.
- Given a data set and the graph of a solution $\mathcal{I}(t)$, we can try to shift and rescale the graph to fit the data.

Change of parameters for SI model solution

- We can rewrite

$$\begin{aligned}\mathcal{I}(t) &= \frac{N}{1 + [N/\mathcal{I}(0) - 1]e^{-pNt}} \\ &= \frac{N}{1 + e^{-\lambda(t-\delta)}} = Ng(\lambda(t-\delta))\end{aligned}$$

  where

$$\begin{aligned}\lambda &= pN \\ \delta &= \log[N/\mathcal{I}(0) - 1]/\lambda \\ g(x) &= 1/(1 + e^{-x}).\end{aligned}$$

- The function g is sometimes called the standard logistic sigmoid function.

Interpretation of new parameters

- If we find parameters N, λ, δ that fit the data, we can solve for the original parameters p and $\mathcal{I}(0)$. However, the new parameters may be more interesting in their own right.
- N is the total number of people who will be infected over the outbreak, according to the model.
- δ is the time at which $N/2$ people have been infected, and at which $d\mathcal{I}/dt$ peaks; it is more relevant than $\mathcal{I}(0)$ to the data and to the intepretation of the model.
- λ is the rate at which the outbreak unfolds; it represents the rate per unit time a single person is infecting others early in the outbreak.

Data Fitting Problems

- Given data points $[t_j, \mathcal{I}_j]$, where $\mathcal{I}_j$ is an estimate of the cumulative number of people infected at time t_j, we can try to minimize the sum of squared residuals

$$E_{\mathcal{I}}(N, \lambda, \delta) = \sum_{j=1}^{J} [\mathcal{I}_j - Ng(\lambda(t_j - \delta))]^2.$$

- If the data is $[t_j, y_j]$ where $t_j = j$ and y_j is the number of new diagnoses per unit time, then we can fit $d\mathcal{I}/dt$ to the data by minimizing

$$E_y(N, \lambda, \delta) = \sum_{j=1}^{J} [y_j - N\lambda g'(\lambda(t_j - \delta))]^2.$$

Partial Solution

- We have posed nonlinear least squares problems.
- Numerical methods for optimization can yield approximate minimizers N, λ, δ.
- We can make some progress algebraically, since E is a quadratic function of N. Minimizing $E_{\mathcal{I}}$ over N yields

$$N_{\lambda,\delta} = \sum_{j=1}^{J} \mathcal{I}_j g(\lambda(t_j - \delta)) \Bigg/ \sum_{j=1}^{J} [g(\lambda(t_j - \delta))]^2.$$

- Substituting and simplifying yields

$$E_{\mathcal{I}}(N_{\lambda,\delta}, \lambda, \delta) = \sum_{j=1}^{J} \mathcal{I}_j^2 - N_{\lambda,\delta} \sum_{j=1}^{J} \mathcal{I}_j g(\lambda(t_j - \delta)).$$

Simple Approaches to Minimizing E

- ("Etch-a-Sketch®") Fix one parameter (say δ) and compute $E(N_{\lambda,\delta}, \lambda, \delta)$ for various λ; look for the value of λ that minimizes E for the chosen value of δ. Then fix λ and adjust δ to make E as small as you can. Then go back and see if you can make E smaller by adjusting λ, etc.
- Make a contour plot of $E(N_{\lambda,\delta}, \lambda, \delta)$ over a range of plausible λ and δ values. Zoom in near the apparent minimum and make another contour plot, etc.
- These approaches can be automated, and of course there are more sophisticated approaches; the latter become important when there are more parameters and/or when the function to be minimized takes a very long time to compute.