Parameter Estimation Lab Session

Marisa Eisenberg (marisae@umich.edu)

Parameter estimation with the SIR model

Let's estimate parameters and investigate the uncertainty in the parameter estimates. We will work with the scaled version of the model, letting $\mu = 0$ since the outbreak data is on a short timescale (so there are probably few births/deaths during this timeframe). The equations are given by:

$$\begin{aligned}
\dot{S} &= -\beta SI \\
\dot{I} &= \beta SI - (\gamma)I \\
\dot{R} &= \gamma I
\end{aligned}$$

where the measurement equation is $y = \kappa I$. The variables S, I, and R represent the proportion of the population that is of susceptible, infectious, and recovered (so $S + I + R = 1$), and we take $\kappa = k \cdot N$, where k is the reporting rate and N is the population size.

1) **Model Simulation**. Write code to simulate the SIR model and plot both the data set (the data set is provided, as `SIRdata.csv`) and the measurement equation $y = \kappa I$. Use the following parameter values: $\beta = 0.4$, $\gamma = 0.25$, $\kappa = 80000$.

For initial conditions, we will let $I(0) = data(0)/\kappa$ (where $data(0)$ is the first data value in `SIRdata.csv`), $S(0) = 1 - I(0)$, and $R(0) = 0$.

2) **Parameter Estimation**. Next, write code to estimate β, γ, and κ using Poisson maximum likelihood and the dataset in `SIRdata.csv`. Use the parameter values in 1) as starting parameter values, and you can use the initial conditions from 1) as well (note though that they depend on κ, which is a fitted parameter—so while we aren't fitting the initial conditions, they will need to change/update as we fit the parameters!). This means you will need to update your initial conditions inside the cost function, so MATLAB/R uses the updated initial conditions when it tries new parameter values.

Plot the data together with your model using the parameter estimates you found. Based on the 'eyeball test', how well does the model fit the data?

3) **Exploring likelihood functions**. Re-run your lab code with some alternative likelihood functions, such as:

- Normally distributed constant measurement error, i.e. ordinary least squares, $Cost = \sum_i (data_i - y_i)^2$.
- Normally distributed measurement error dependent on the data, weighted least squares, e.g. using Poisson-style variance, $Cost = \sum_i \frac{(data_i - y_i)^2}{data_i}$. This assumes the variance at any given data point is equal to the data, but you can also try other weightings!
- Extended/weighted least squares, e.g. also using Poisson-style variance, $Cost = \sum_i \frac{(data_i - y_i)^2}{y_i}$. This assumes the variance at any given data point i is equal to y_i, the model prediction at that time.
- Maximum likelihood assuming other distributions for the observation error, e.g. negative binomial

How do the parameter estimates and/or uncertainty differ from the estimates you got earlier? What are the underlying assumptions for on the model/measurement equation/likelihood that generate the different least squares cost functions given above?

Extra problem: The recovery rate γ is often approximately known, so let's fix the value of $\gamma = 0.25$. Now we have two unknown parameters, β and κ. Plot the likelihood as a surface or heat map as a function of β and κ (i.e. so that color is the likelihood value, and your x and y axes are the β and κ values respectively. How does the shape of the likelihood change as you switch likelihood functions?

4) **Un-scaled SIR model**. Try out estimating the parameters and evaluating uncertainty for the un-scaled SIR model given as:

$$\begin{aligned} \dot{S} &= \mu N - bSI \\ \dot{I} &= bSI - \gamma I \\ \dot{R} &= \gamma I \end{aligned}$$

How does the optimizer perform? How do the profile likelihoods and rank of the FIM turn out? Note that to fit this model you will need to also fit the initial conditions (or at least $S(0)$), since we no longer can say that $S + I + R = 1$, instead $S + I + R = N$, where N is the unknown total population.

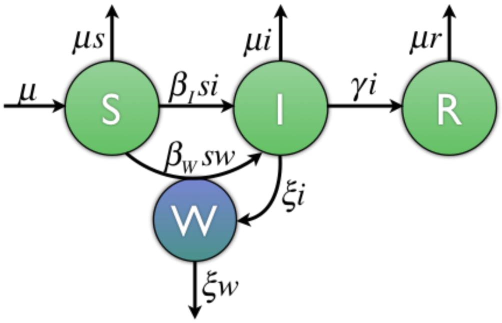

Figure 1: SIWR model of cholera transmission.

5) **Cholera Transmission in Yemen**. Cholera and many waterborne diseases exhibit multiple pathways of infection, which can be modeled (for example) as direct and indirect transmission. A major public health issue for waterborne diseases involves understanding the modes of transmission in order to improve control and prevention strategies (see e.g. Hartley 2006). An important epidemiological question is therefore: given data for an outbreak, can we determine the role and relative importance of direct (human-mediated) vs. environmental/waterborne routes of transmission?

To examine this question, we will use the SIWR model developed by Tien and Earn (2010), shown in Figure 1. We will combine this model with the data on deaths over time from the ongoing cholera epidemic in Yemen (data is from the Humanitarian Data Exchange). The scaled SIWR model is given by the following equations:

$$\begin{aligned}\dot{S} &= \mu - \beta_I SI - \beta_W SW - \mu S\\ \dot{I} &= \beta_I SI + \beta_W SW - (\mu + \gamma)I\\ \dot{W} &= \xi(I - W)\\ \dot{R} &= \gamma I - \mu R\end{aligned}$$

where

- S, I, and R are the fractions of the population who are susceptible, infectious, and recovered
- W is a scaled version of the concentration of bacteria in the water
- β_I and β_W are the transmission parameters for direct (human-human) and indirect (environmental) cholera transmission
- ξ is the pathogen decay rate in the water
- γ is the recovery rate
- μ is the birth/death parameter for the population

Since we are considering a short-term outbreak (less than one year), it is reasonable to assume that the effects of births and deaths are negligible, so we set $\mu = 0$. In addition, the recovery time for cholera is reasonably well known, so we can fix $\gamma = 0.25$ based on previous work (Tuite 2011, etc.) (i.e. we don't need to estimate this). The SIWR model has previously been shown to be structurally identifiable using the differential algebra approach (Eisenberg 2013).