Mathematical Cell Biology

Suggested data-fitting problems

May 2012

1. **Predicting housing prices using multiple linear regression:** The table below contains the sales prices of 5 holiday cottages in Odsherred, Denmark, together with the age and the livable area of each cottage
(source - `http://statmaster.sdu.dk/courses/st111/index.html`)

Price DDK 1000	Age years	Area m^2
745	36	66
895	37	68
442	47	64
440	32	53
1598	1	101

We will fit this data to a linear model of the form

$$\text{price} = a + b \cdot \text{age} + c \cdot \text{area}$$

(a) Construct the design matrix (also called the Vandermonde matrix in Matlab documentation) for the given data and the linear regression model above.

(b) Use the Matlab backslash operator, \ to solve for the best-fit values of the parameters a, b and c.

(c) Plot the residuals, and calculate the r^2 for the fit.

2. **Enzyme kinetics:** Recall the Michelis-Menten equation for the rate of an enymatic reaction, v, for a given substrate concentration, S:

$$v = \frac{v_{\text{max}} S}{K_{\text{M}} + S}$$

The following table lists the speed of an enzyme-catalyzed reation for various susbstrate concentrations and under three different conditions, V_1 with the enzyme alone, and V_2 and V_3 with two different inhibitors, respectively.

S	V_1	V_2	V_3
1	0.191	0.153	0.106
5	0.414	0.364	0.221
15	0.484	0.471	0.274
30	0.521	0.521	0.287
45	0.514	0.516	0.296

(a) While this is not a recommended procedure for data-fitting, the above equation can be transformed to a linear relationship between $1/v$ and $1/S$ (Lineweaver-Burk plot). In the lecture, we derived exact analytical expressions for best-fit parameter estimated for a linear fit. Use these results to calculate the best-fit values of v_{max} and K_{M} for each of the three cases.

(b) Next, fit the data to a straight line using Matlab and check whether the results match those from part (a).

(c) Plot all three linear fits and classify the two inhibitors (competitive or non-competitive).

(d) Finally, use the Matlab function **nlinfit** to fit the data to the original nonlinear form of the Michelis-Menten equation and compare the best-fit parameter estimates to those from parts (a) and (b).

(e) Plot the data and the nonlinear regression curves for all three cases.

3. **Fitting FRAP data:** Download the FRAP recovery data in the attached text files: (`integrin_YFP_untreated.dat` and `integrin_YFP_cpz.dat`). The first column in each file contains time points, and the remaining columns show the fractional recovery $f(t)$ in a FRAP experiment with fluorescently labelled integrins. Each column is from an independent replicate experiment.

   (a) For each dataset, ***simultaneously*** fit all the recovery curves to a two parameter recovery model:

   $$f(t) = f_{\max}\left(1 - e^{t/\tau}\right)$$

   and estimate the model parameters $f_{\max}$ and τ.

   (b) Recall that these model parameters are related to the kinetic rate constants k_{on} and k_{off} for a 2-state model of integrin turnover. Using the fitted values of the parameters above, estimate the two rate constants for each dataset.

4. **Model selection for dose-response data** Download the dose-response data in the attached file (`DoseResponse.dat`). The first column is $x = \log_{10}(\text{dose})$ and the second column is the response y.

   (a) Fit the data with two alternate models: a fixed-slope Hill function:

   $$y = \frac{E_{\max}}{1 + 10^{(\mathrm{LEC}_{50} - x)}}$$

   where $\mathrm{LEC}_{50} = \log_{10}(\mathrm{EC}_{50})$, or a standard Hill function:

   $$y = \frac{E_{\max}}{1 + 10^{n(\mathrm{LEC}_{50} - x)}}$$

   where n is the Hill coefficient. For each model estimate the best-fit values of the model parameters.

   (b) Are these two models nested? If so, use the F-test to establish which model is a better descriptor of the data.

   (c) Calculate Akaike's information criterion for each fit, and assign weights to the two models based on their AIC value. Use the second order corrected expression for AIC:

   $$\mathrm{AIC_c} = n \log\left(\frac{\mathrm{SSR}}{n}\right) + 2k + \frac{2k(k+1)}{n-k-1}$$

5. *[Advanced]* **Fitting data to an epidemic model:** In this problem we consider an elementary SI model for an epidemic that is described by the following system of coupled nonlinear ODE's:

   $$\begin{aligned} dS/dt &= \lambda - \delta_1 S - rSI \qquad && S(0) = S_0 \\ dI/dt &= rSI - \delta_2 I && I(0) = I_0 \end{aligned}$$

where S and I are the susceptible and infected subsets of the population, λ is the birth rate, δ_1 and δ_2 are the per-capita mortality rates of the susceptible and infected populations, respectively, and r is the transmission rate. Note that this model assumes no recovery from the infection (that is, once infected, an individual stays that way).

Download the attached file `Epidemic.dat` that contains simulated data for an epidemic. The three columns are t, S and I, respectively. This example is fundamentally different from the other data-fitting problems we have looked at, in that, there are now two dependent variables (S and I) for a single independent variable t.

(a) Write a Matlab function that will numerically integrate the above ODE system for input parameter values and initial conditions, and output the population in the two classes at the observation times. Use the built-in ODE solver **ode45**.

(b) Next, write a function that calculates the sum of squares for the given data and user-supplied parameter values.

(c) Minimize the sum of squares with respect to the model parameters starting with some initial guesses for the parameter values. Try both the functions **fminuncon** and **fmincon**. What are some reasonable constraints on the parameters?

(d) Can the parameters be optimized using the function **nlinfit**? If so, how well do the parameter estimates match those calculated in part (c)?