BME 2200: Biostatistics and Research Methods

Lecture 8: Least-squares curve fit and non-linear regression

Lucas C. Parra
Biomedical Engineering Department
City College of New York

parra@ccny.cuny.edu

Content, Schedule

1. Scientific literature:

- Literature search
- Structure biomedical papers, engineering papers, technical reports
- Experimental design, correlation, causality.

2. Presentation skills:

- Report – Written report on literature search (individual)
- Talk – Oral presentation on biomedical implant (individual and group)

3. Graphical representation of data:

- Introduction to MATLAB
- Plot formats: line, scatter, polar, surface, contour, bar-graph, error bars. etc.
- Labeling: title, label, grid, legend, etc.
- Statistics: histogram, percentile, mean, variance, standard error, box plot

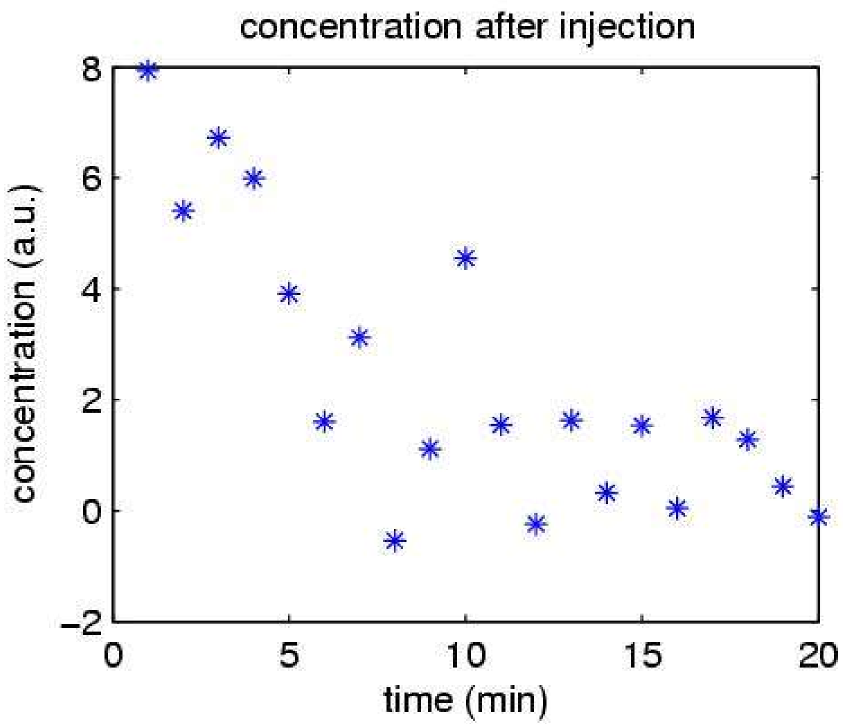
4. Biostatistics:

- Basics of probability
- t-Test, ANOVA
- Linear regression, Least-squares curve fit
- Error analysis
- Test power, sensitivity, specificity, ROC analysis

Non-linear dependencies

Sometime one may want to capture a non-linear dependence between two variables.

For instance, assume that we inject a drug into the blood stream, and want to quantify how long the drug takes to clear from the blood. We may measure a blood plasma concentration of the drug at some regular intervals and observe something like this:



Example: exponential decay

A reasonable model would be to assume that the amount of clearance of the substance is proportional to its concentration

$$\tau \frac{dc}{dt} = -c$$

This very simple differential equation is solved by $c(t) = a\, e^{-t/\tau}$ which describes an exponentially decaying concentration.

Where $a=c(0)$ is the concentration at the beginning $t=0$, and τ is the time it takes for the concentration to decay to $1/e$ of its starting value.

We now have to find these parameters from our measurements

$$y_i = c(t_i) \qquad x_i = t_i$$

Non-linear least-squares fit

The goal is therefore to find the parameters a and τ that best reproduce the observation y_i from x_i:

We generate estimate $$\hat{y}_i = a\, e^{-x_i/\tau} = f(x_i; a, \tau)$$

The least squares estimate for the parameters is then

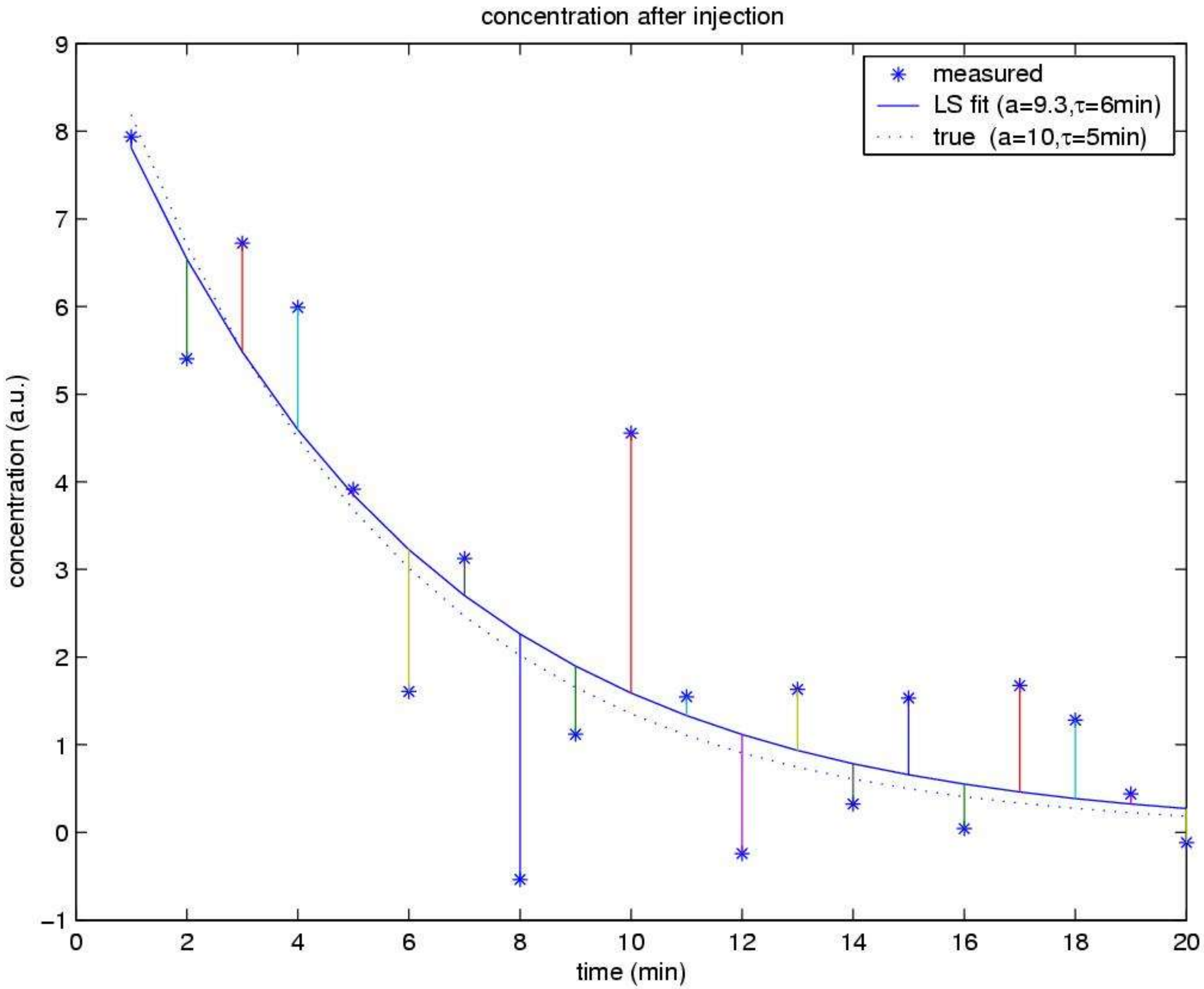
$$a, \tau = \underset{a,\tau}{argmin} \sum_i (y_i - \hat{y}_i)^2$$

This optimization problem can not be solved analytically and must be solved numerically. In matlab we can use the function `lsqcurvevefit()`:

```
>> fun = inline('p(1)*exp(-x/p(2))','p','x');
>> p = lsqcurvefit(fun,p0, x, y)
```

Example: Exponential decay

For the data shown before we obtain



Assignment:

Assignment 9:

Reproduce figure on slide #6:

1. Generate concentration data that follows an exponentially decaying trend.

2. Add noise according to the corresponding model.

3. Use lsqcurvefit() to find the model parameter that will give the optimal fit to the data.

4. Computed the estimated concentrations with the optimal parameters.

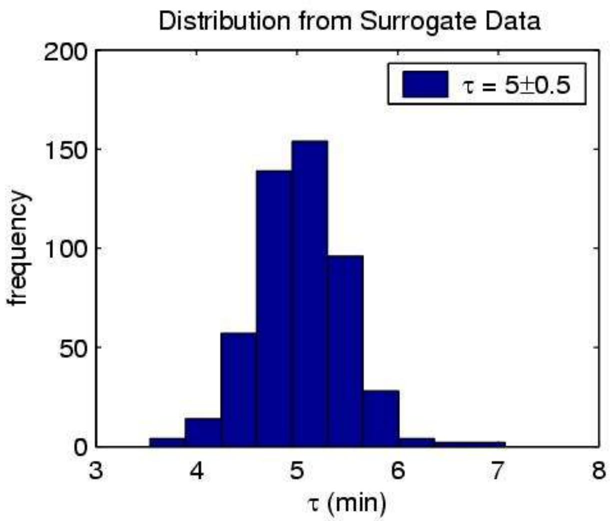
5. Display them together with the original concentrations and the noisy version.

Estimation accuracy with surrogate data

Note that we could have used any other non-linear relationship. The 'correct' choice is determined by our prior knowledge of the problem.

Assuming that the choice is correct one would like to know how certain we are about the estimated parameters τ and a. The analytic solutions to this question is complicated.

An *approximate numerical solutions* can be obtained by generating ***surrogate data sets*** and repeating the fitting procedure on each surrogate. This gives an distribution for each parameter, and we can give the standard deviation of these solutions as approximate error bars.



Transform non-linear to linear problem

We have assumed here that the noise (or the error of the model) is in the measurement $c(t_i)$ and minimized the error

$$e_i = y_i - \hat{y}_i$$

If the error is instead in t_i then we should define $y_i = t_i$

For the exponential model we are lead back to a linear regression problem

$$t = \tau \log c(t) - \tau \log a$$

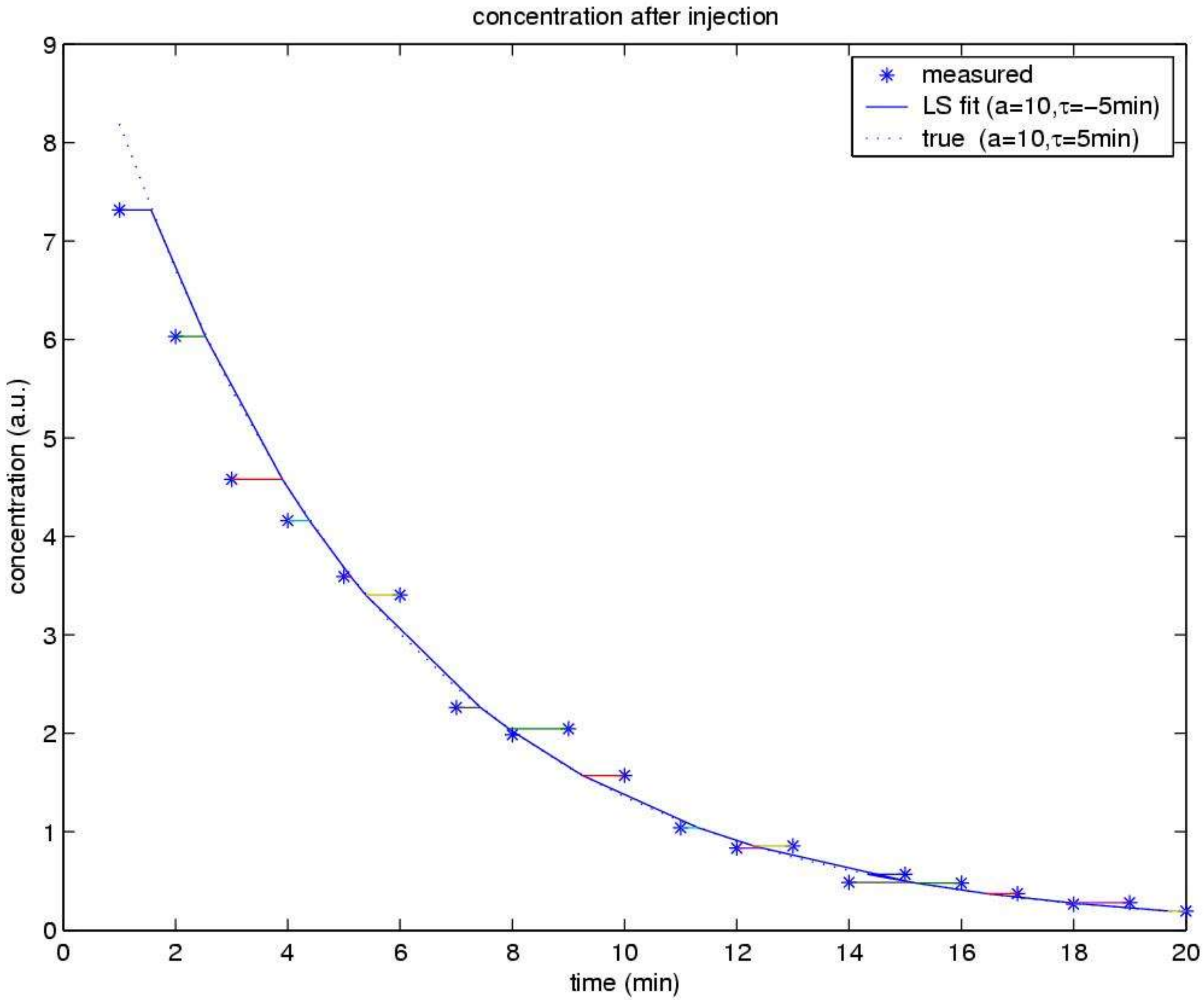
If we define $x_i = \log c(t_i)$ and substitute $\tau \rightarrow b, -\tau \log a \rightarrow a$ as new parameters to be estimated

$$y_i = b x_i + a$$

This trick of transforming a non-linear problem to a linear problem only works in a few special cases.

Transform non-linear to linear problem

Really only legitimate if the error is in the time *t*.



Non-linear regression

Another approach to reduce the regression of a non-linear problem to a linear problem is to use a set of basis functions.

$$y = \sum_j a_j f_j(x)$$

The functions can be for instance

$$f_0(x) = 1, f_1(x) = x, f_2(x) = x^{2,} \ldots, f_n(x) = x^n$$

The solution for the vector $\boldsymbol{a}=[a_1, \ldots, a_n]^T$ is given by

$$\boldsymbol{a} = (\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}\, \boldsymbol{y}$$

where ij-th element of the matrix $\boldsymbol{F}$ is $f_j(x_i)$ and $\boldsymbol{y}=[y_1, \ldots, y_n]^T$

Non-linear regression

Example: Fitting a parabola ... in class