BSR 1803
Systems Biology: Biomedical Modeling

Model Fitting and Error Estimation

Kevin D. Costa
Steven Kleinstein and Uri Hershberg

Spring 2010

Biomathematical Model

- A system of mathematical equations or computer simulations that provides a quantitative picture of how a complex biological system functions under healthy and diseased conditions.

- Computational models use numerical methods to examine mathematical equations or systems of equations too complex for analytical solution.

Advantages of the Modeling Approach

- Concise summary of present knowledge of operation of a particular system
- Predict outcomes of modes of operation not easily studied experimentally in a living system
- Provide diagnostic tools to test theories about the site of suspected pathology or effect of drug treatment
- Clarify / simplify complex experimental data
- Suggest new experiments to advance understanding of a system

Limitations of the Modeling Approach

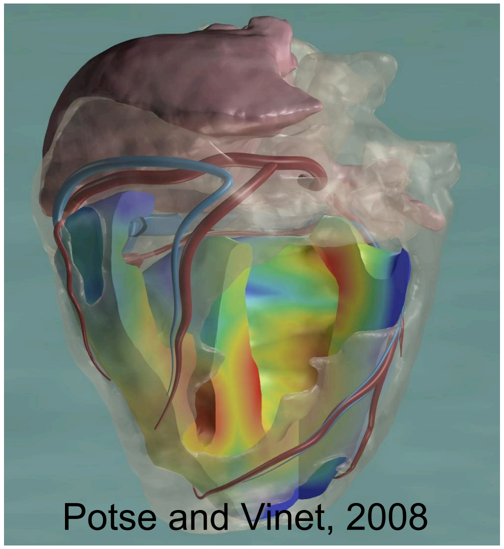
- Models often require many simplifying assumptions
  - garbage in, garbage out
- Validation of model predictions is essential
  - examination of behavior under known limiting conditions
  - experimental validation
  - limits of model point out what we don’t understand

Perspectives to Keep in Mind

"What we observe is not nature in itself but nature exposed to our method of questioning." W. Heisenberg

"Any model is only ever a model--experiments are the truth!" J.W. Covell

Forward Model


Potse and Vinet, 2008

- A detailed mathematical model designed to incorporate a desired level of anatomic, physical, or physiologic features
  - Can have arbitrary complexity as desired
  - Parameter values often obtained from published literature
  - Ex: cardiac electromechanical coupling, cell signaling networks
- Used for simulating realistic experimental data under precisely defined conditions to test hypotheses *in silico*
- Can help design better experiments and reduce animal use
- Generally too complicated for fitting to experimental data
- Allows generation of synthetic data sets with prescribed noise characteristics (Monte Carlo simulation) for evaluating parameters obtained by inverse modeling

Inverse Model

- A mathematical model designed to fit experimental data so as to explicitly quantify physical or physiological parameters of interest
- Values of model elements are obtained using parameter estimation techniques aimed at providing a “best fit” to the data
- Generally involves an iterative process to minimize the average difference between the model and the data
- Evaluating the quality of an inverse model involves a combination of established mathematical techniques as well as intuition and creative insight

Forward-Inverse Modeling

- A process of combined data simulation and model fitting used for evaluating the robustness, uniqueness, and sensitivity of parameters obtained from an inverse model of interest.

- A powerful tool for improving data analysis and understanding the limitations on model parameters used for system characterization and distinguishing normal from abnormal populations.

Characteristics of a Good Inverse Model

- Fit is good—model should be able to adequately describe a relatively noise-free data set (of course a poor fit provides some insight also).
- Model parameters are unique
  - Theoretically identifiable for noise-free data
  - Well-determined model parameters in presence of measurement noise
- Values of parameter estimates are consistent with hypothesized physical/physiologic meanings and change appropriately in response to alterations in the physiologic system.

Steps for Inverse-Modeling of Data

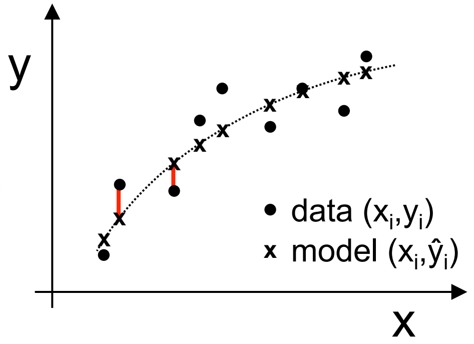
1. Select an appropriate mathematical model
   - Polynomial or other functional form
   - Based on underlying theory
2. Define a “figure of merit” function
   - Measures agreement between data & model for given parameters
3. Adjust model parameters to get a “best fit”
   - Often involves minimizing the figure of merit function
4. Evaluate “goodness of fit” to data
   - Never perfect due to measurement noise
5. Estimate accuracy of best-fit parameter values
   - Provide confidence limits and determine uniqueness
6. Determine whether a much better fit is possible
   - Tricky due to possible local minima vs global minimum
   - F-test for comparing models of different complexity

Selecting the Model

- "Trend lines"
  - Polynomials are often used when a data set seems to follow a mathematical trend but the governing formula is not known

- Physically-based equations
  - Given knowledge of a governing physical process, the desired model is derived from the underlying theoretical equations
  - Resulting model parameters have a specific physical interpretation

Least-Squares Error Minimization



- Goal is to fit N data points (x_i, y_i) i=1..N
- The model is a function with M adjustable parameters (degrees of freedom) a_k, k=1..M used to generate N model points (x_i, $\hat{y}_i$)

$$\hat{y}_i = \hat{y}(x_i, a_1 .. a_M)$$

- The residual measures the difference between a data point and the corresponding model estimate

$$y_i - \hat{y}(x_i, a_1 .. a_M)$$

- Since residuals can be positive or negative, a sum of residuals is not a good measure of overall error in the fit

$$\sum_{i=1}^{N} [y_i - \hat{y}(x_i, a_1 .. a_M)]$$

- A better measure is the sum of squared residuals, *E*, which is only zero if every residual is zero

$$E = \sum_{i=1}^{N} [y_i - \hat{y}(x_i, a_1 .. a_M)]^2$$

Maximum Likelihood Estimation

- Not meaningful to ask “What is the probability that my set of model parameters is correct?”
  - Only one correct parameter set—Mother Nature!
- Better to ask “Given my set of model parameters, what is the probability that this data set occurred?”
  - What is the likelihood of the parameters given the data?
- Inverse modeling is also known as “maximum likelihood estimation”.

The Chi-Square Error Measure and Maximum Likelihood Estimation

- For Gaussian distribution of measurement noise with varying standard deviation, σ_i, the probability of the data set coming from the model parameters is given by

$$P \propto \prod_{i=1}^{N} \exp\left(-\frac{[y_i - \hat{y}(x_i)]^2}{2\sigma_i^2} \right)$$

- Maximizing this probability involves maximizing ln(P) or minimizing –ln(P), yielding the chi-square function of weighted residuals
  - the “weight” is the inverse of the variance of each measurement ($w_i = \sigma_i^{-2}$)
  - Other functions may be useful for non-Gaussian measurement noise, yielding so-called “robust estimation” methods

$$-\ln(P) \propto \sum_{i=1}^{N} \frac{[y_i - \hat{y}(x_i)]^2}{\sigma_i^2} \equiv \chi^2$$

- If variance is assumed to be uniform, then let σ = constant = 1, and chi-square function yields the sum of squared residuals function defined earlier

$$\chi^2 \Big|_{\sigma=1} = \sum_{i=1}^{N} [y_i - \hat{y}(x_i)]^2 = E$$

Minimizing Chi-Square

- Since the error in the model fit depends on the model parameters, a$_k$, minimizing the chi-square function requires finding where the derivatives are zero

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - \hat{y}(x_i)]^2}{\sigma_i^2}$$

$$\frac{\partial(\chi^2)}{\partial a_k} = -2\sum_{i=1}^{N}\left(\frac{[y_i - \hat{y}(x_i)]}{\sigma_i^2}\right)\left(\frac{\partial \hat{y}(x_i, a_1..a_M)}{\partial a_k}\right) = 0\,; \quad k = 1..M$$

- This yields a general set of M (nonlinear) equations for the M unknowns a$_k$
- The model derivatives dŷ/da$_k$ are often known exactly, or may be approximated numerically using finite differences

Linear Regression Analysis

- Consider a set of measurements of photodetector voltage (dependent variable) as a function of incident laser intensity (independent variable).

- We propose to examine a linear relationship between voltage (*y*) and intensity (*x*).

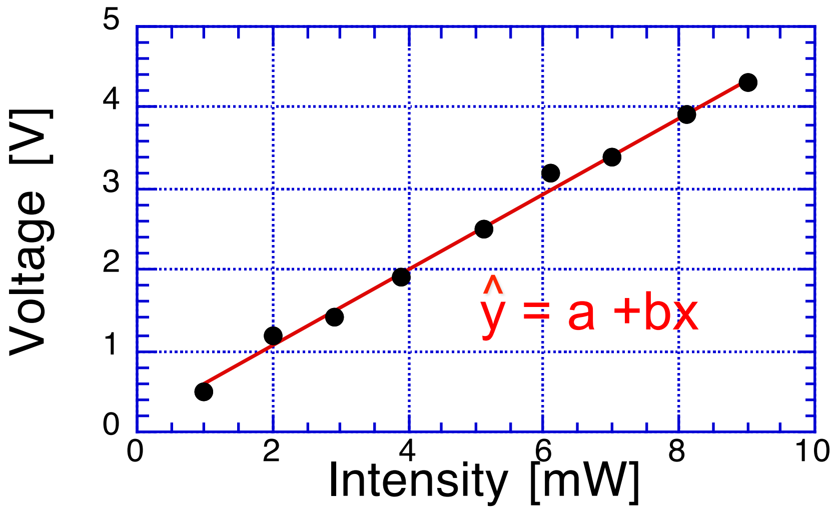
$$\hat{y} = a + bx$$

data $x = [x_1, x_2, \ldots, x_n]$

data $y = [y_1, y_2, \ldots, y_n]$

model $\hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n]$

- We need to find the best set of values of *a* and *b* to fit the data.



- Define the least-squares error norm defining the “goodness” of the linear fit. Adjust model parameters *a* and *b* to minimize this error.

$$E(a,b) = \sum_{i=1}^{N} \left[y_i - (a + bx_i) \right]^2$$

Computing Model Parameters for Linear Regression

- We can determine the best values of *a* and *b* by calculating the partial derivatives of *E* w.r.t. *a* and *b*, and setting these to zero. This yields 2 equations to be solved for the 2 unknowns *a* and *b*, yielding:

$$E(a,b) = \sum_{i=1}^{N}\left[y_i - (a + bx_i)\right]^2$$

$$\frac{\partial E(a,b)}{a} = 0 \qquad \frac{\partial E(a,b)}{b} = 0$$

$$b = \frac{\Sigma x_i y_i - n\,\bar{x}\,\bar{y}}{\Sigma x_i^2 - n(\bar{x})^2} \qquad a = \bar{y} - b\,\bar{x}$$

- Standard error of the estimate approximates standard deviation of population about mean at a given value of the independent variable

$$s_{x,y} = \sqrt{\frac{n-1}{n-2}(s_y^2 - b^2 s_x^2)}$$

- Standard error of slope and intercept used for *t* test of *a*,*b* = 0 or to place confidence intervals

$Y = \alpha + \beta X + \epsilon$

$$s_a = s_{x,y}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

$$s_b = \frac{1}{\sqrt{n-1}}\,\frac{s_{x,y}}{s_x}$$

Regression versus Correlation

- Correlation coefficient describes the strength of the association between the two variables
  - r → +1 if they increase together
  - r → -1 if one decreases as other increases
  - r → 0 if they do not relate to one another
- The correlation coefficient can be related to results of the regression
- Unlike the regression parameters, *a* and *b*, the correlation coefficient, *r*, is symmetric in *x* and *y* and therefore does not require choosing of independent and dependent variables

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

$$= \sqrt{1 - \frac{(n-2)}{(n-1)} \frac{s_{x,y}^2}{s_y^2}}$$

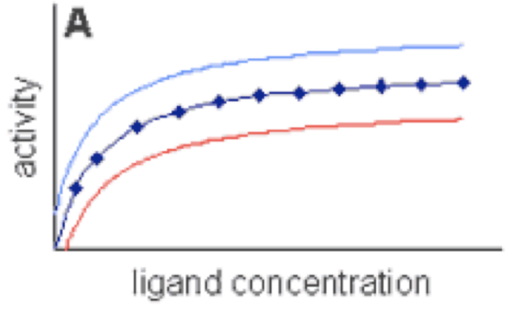
$$b = \frac{\Sigma x_i y_i - n\,\bar{x}\,\bar{y}}{\Sigma x_i^2 - n(\bar{x})^2}$$

$$a = \bar{y} - b\,\bar{x}$$

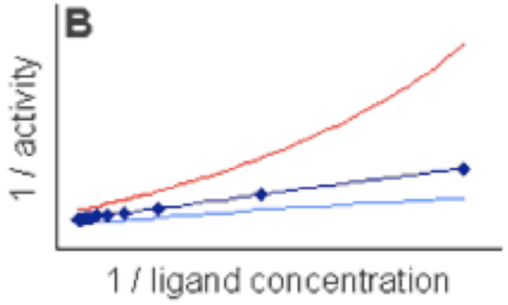
Linearization of Nonlinear Models

- Many nonlinear equations can be "linearized" by selecting a suitable change of variables
- Historically this has been a common approach in analysis of scientific data, mainly due to ease of implementation
- However, "linearization" often distorts the error structure, violates key assumptions, and impacts resulting model parameter values, which may lead to incorrect conclusions
- In our modern era of computers it is usually wisest to perform nonlinear least squares analysis when using nonlinear inverse models

$$V = V_{\max} \frac{[S]}{K_m + [S]}$$



$$\frac{1}{V} = \frac{K_m}{V_{\max}} \frac{1}{[S]} + \frac{1}{V_{\max}}$$



adapted from Lobemeier, 2000

General Model Fitting

- It is important to understand where these regression equations come from, but this is rarely done by hand.
- Microsoft Excel has several trend-line functions built in, including nonlinear models which follow the same idea but cannot be solved analytically.
- Often in biomedical experiments, a data set is governed by a system of equations determined by underlying physical principles rather than just the shape of the curve.

Nonlinear Model Fitting

- The selected model ŷ is a nonlinear function of model parameters a_k, k=1..M

$$\hat{y}_i = \hat{y}(x_i, \mathbf{a})$$

- The χ^2 merit function is

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N} \frac{[y_i - \hat{y}(x_i, \mathbf{a})]^2}{\sigma_i^2}$$

- The gradients of χ^2 with respect to model parameters a_k must approach zero at minimum χ^2

$$\frac{\partial(\chi^2)}{\partial a_k} = -2\sum_{i=1}^{N}\left(\frac{[y_i - \hat{y}(x_i, \mathbf{a})]}{\sigma_i^2}\right)\left(\frac{\partial \hat{y}(x_i, \mathbf{a})}{\partial a_k}\right)$$

- However, because the gradients are nonlinear functions of **a**, minimization must proceed iteratively updating **a** until χ^2 stops decreasing.

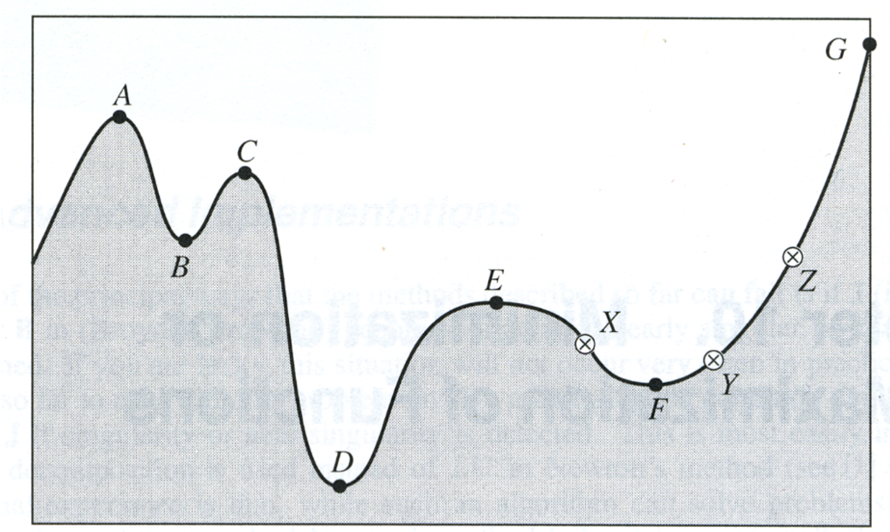
- In the steepest descent method, the constant, λ, must be small enough not to exhaust the downhill direction.

$$\mathbf{a}_{next} = \mathbf{a}_{current} - \lambda \times \nabla\chi^2(\mathbf{a}_{current})$$

- Alternative numerical methods include the inverse-Hessian method, the popular hybrid Levenberg-Marquardt method, and the robust but complex full Newton-type methods.

Global Error Minimization

- The error function depends on model parameters a_k, and can be thought of as an M-dimensional "surface" of which we seek the minimum
- Depending on the complexity of the model (i.e. the number of model degrees of freedom, M) the error surface may be quite "bumpy"



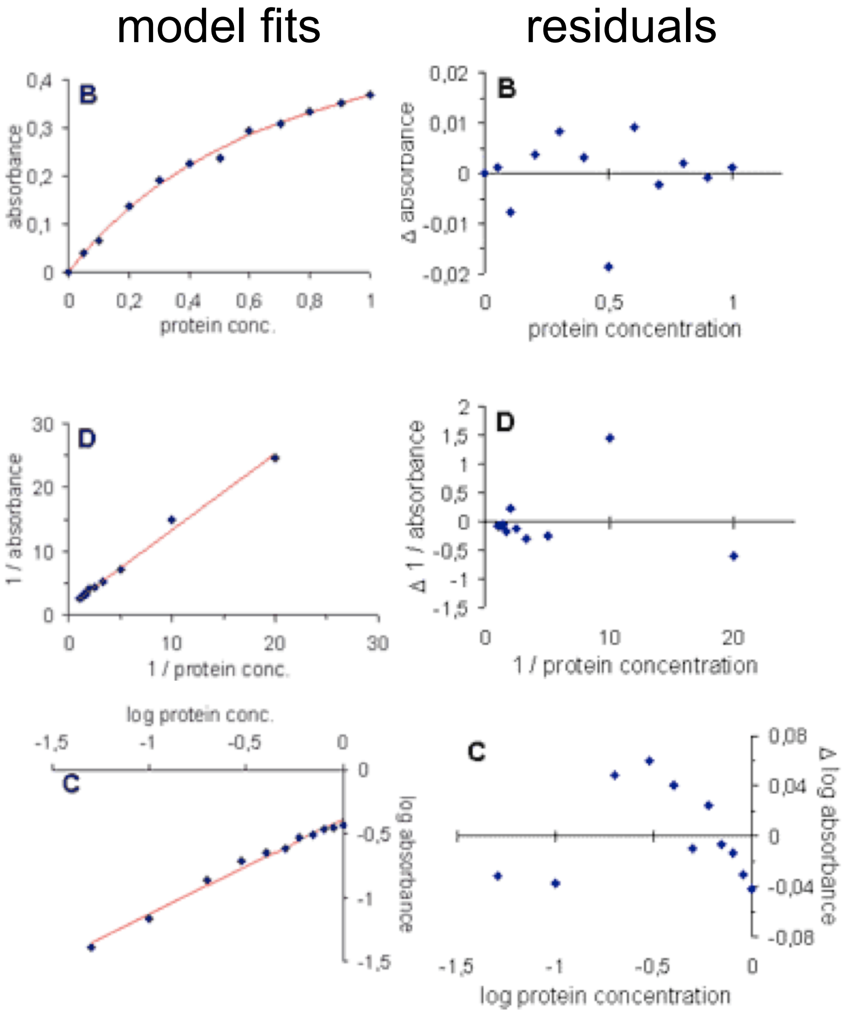
- A challenge is to ensure that a given set of "optimal" model parameters represents the true global minimum of the error surface, and not a local minimum
- This can be tested by varying the initial guesses and comparing the resulting model parameters

Implementation in Matlab

```
function KDC_optimization
global known;
filename = input('Enter the name of file: ','s');
data = dlmread(filename);
x_data = data(:,1);
y_data = data(:,2);
known = 10;                             % Assign known model parameters
guess = [.1 .1 1 1];                    % Guess initial values
[optimum,resnorm] = lsqnonlin(@model,guess,LB,UB,options,x_data,y_data)

y_model=model(optimum,x_data);          % Generate vector of simulated data

plot(x_data,y_data,'bx',x_data,y_model,'r-');
xlabel('Independent Variable (***)');
ylabel('Dependent Variable (***)');

function y=model(a,x)
global known;
y=a(1)+a(2)*x.^2+a(3).*sin(a(4).*x) - known;    % May depend on known variables
```

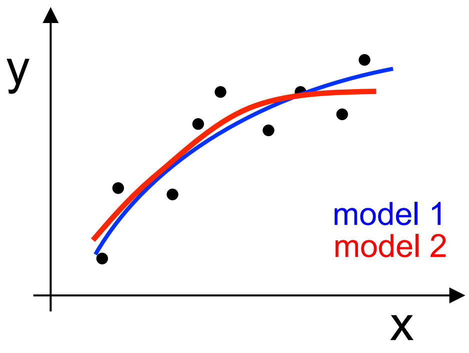
Goodness of Fit and the Residuals Plot

- The correlation coefficient (R^2) is often used to characterize the goodness of fit between model and data.
- A high correlation can exist even for a model that systematically differs from the data.
- One must also examine the distribution of residuals--a good model fit should yield residuals equally distributed along x and normally distributed around zero with no systematic trends



adapted from Lobemeier, 2000

Comparing Two Model Fits



- The number of data points, N, must exceed the number of model parameters, M, yielding the degrees of freedom (DOF = N-M)

$$M \le N - 1$$

- Increasing the number of model parameters, M, will generally improve the quality of fit and reduce χ^2
- The mean squared error can be used to compare two models fit to a given data set

$$MSE = \frac{\chi^2}{N-M} = \frac{\chi^2}{DOF}$$

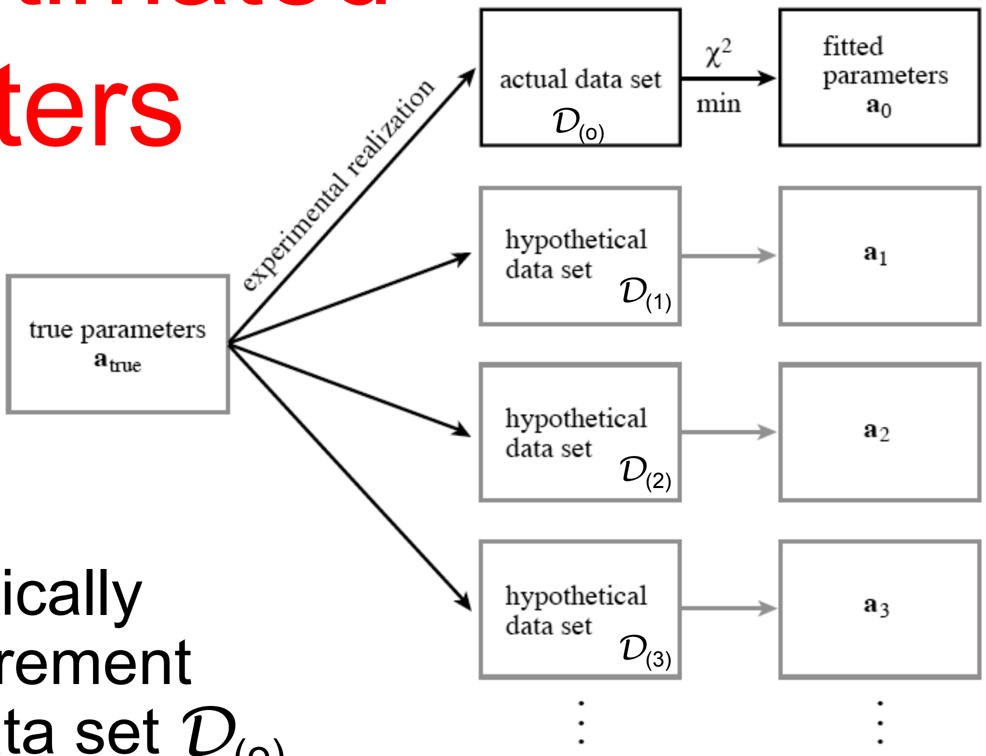
- Increasing MSE with decreasing χ^2 can reveal an over-parameterized model
- An F-statistic can be computed for the results of two model fits.
  - F~1, the simpler model is adequate
  - F > 1, the more complex model is better, or random error led to a better fit with the complex model
  - P-value defines the probability of such a "false positive" result

$$F = \frac{\dfrac{(\chi^2_{simple} - \chi^2_{complex})}{(DOF_{simple} - DOF_{complex})}}{\dfrac{\chi^2_{complex}}{DOF_{complex}}}$$

Table 3–1 Critical Values of *F* Corresponding to $P < .05$ (Lightface) and $P < .01$ (Boldface)

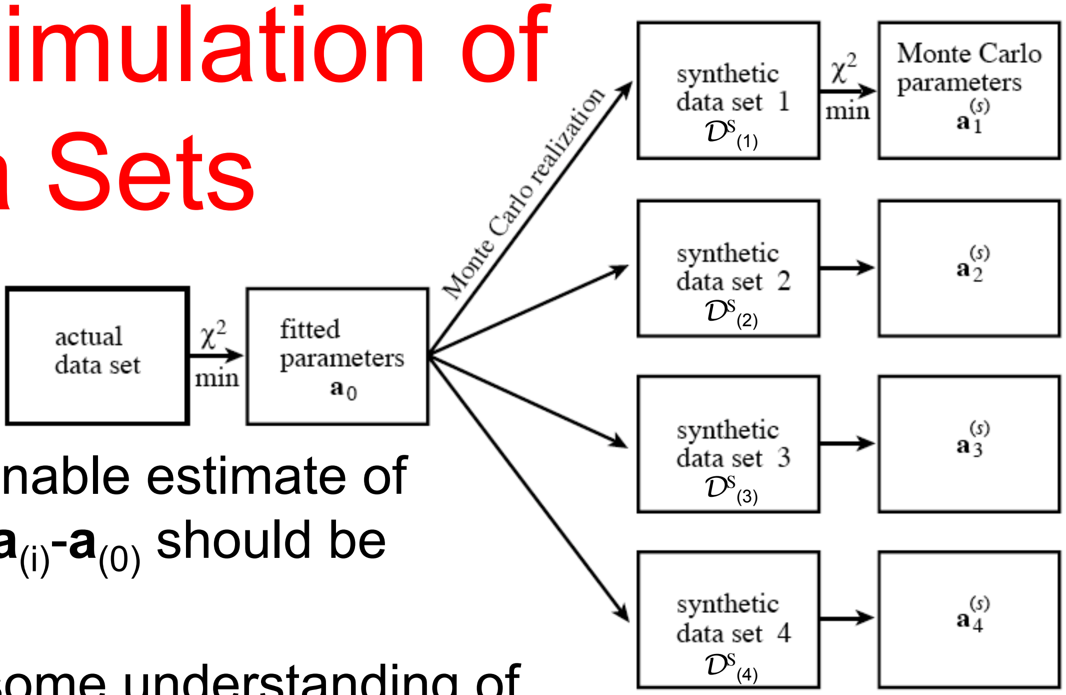
	ν_n																							
ν_d	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4052	**4999**	**5403**	**5625**	**5764**	**5859**	**5928**	**5981**	**6022**	**6056**	**6082**	**6106**	**6142**	**6169**	**6208**	**6234**	**6261**	**6286**	**6302**	**6323**	**6334**	**6352**	**6361**	**6366**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.36**	**99.37**	**99.39**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.29**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**

Accuracy of Estimated Model Parameters



- Underlying true set of model parameters, $\mathbf{a}_{\text{true}}$, are known to Mother Nature but hidden from the experimenter
- True parameters are statistically realized, along with measurement errors, as the measured data set $\mathcal{D}_{(o)}$
- Fitting $\mathcal{D}_{(o)}$ using χ^2 minimization yields the estimated model parameters $\mathbf{a}_{(o)}$
- Other experiments could have resulted in data sets $\mathcal{D}_{(1)}$, $\mathcal{D}_{(2)}$, etc. which would have yielded model parameters $\mathbf{a}_{(1)}$, $\mathbf{a}_{(2)}$, etc.
- We wish to estimate the probability distribution of $\mathbf{a}_{(i)}$ - $\mathbf{a}_{\text{true}}$ without knowing $\mathbf{a}_{\text{true}}$ and without an infinite number of hypothetical data sets. Hmmmm…

Monte Carlo Simulation of Synthetic Data Sets



actual data set

χ^2 min

fitted parameters $\mathbf{a}_0$

Monte Carlo realization

synthetic data set 1 $\mathcal{D}^S_{(1)}$

χ^2 min

Monte Carlo parameters $\mathbf{a}_1^{(s)}$

synthetic data set 2 $\mathcal{D}^S_{(2)}$

$\mathbf{a}_2^{(s)}$

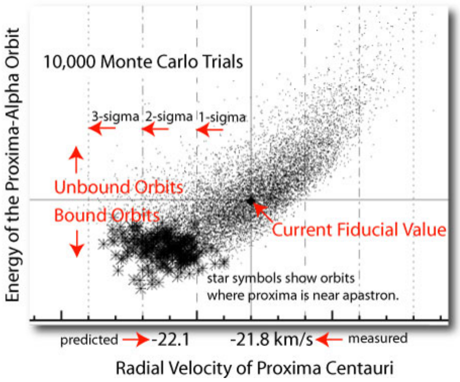
synthetic data set 3 $\mathcal{D}^S_{(3)}$

$\mathbf{a}_3^{(s)}$

synthetic data set 4 $\mathcal{D}^S_{(4)}$

$\mathbf{a}_4^{(s)}$

- Assume that if $\mathbf{a}_{(0)}$ is a reasonable estimate of $\mathbf{a}_{true}$, then the distribution of $\mathbf{a}_{(i)}$-$\mathbf{a}_{(0)}$ should be similar to that of $\mathbf{a}_{(i)}$-$\mathbf{a}_{true}$
- With the assumed $\mathbf{a}_{(0)}$, and some understanding of the characteristics of the measurement noise, we can generate "synthetic data sets" $\mathcal{D}^S_{(1)}$, $\mathcal{D}^S_{(2)}$,… at the same x_i values as the actual data set, $\mathcal{D}_{(o)}$, have the same relationship to $\mathbf{a}_{(0)}$ as $\mathcal{D}_{(o)}$ has to $\mathbf{a}_{true}$.
- For each $\mathcal{D}^S_{(1)}$, perform a model fit to obtain corresponding $\mathbf{a}^S_{(j)}$, yielding one point $\mathbf{a}^S_{(j)}$- $\mathbf{a}_{(0)}$ for simulating the desired M-dimensional probability distribution. **This is a very powerful technique!!**
- Note: if σ_i^2 are not known, can estimate after fit and use `randn` function in Matlab



$$\sigma^2 = \frac{\sum_{i=1}^{N}[y_i - \hat{y}(x_i,\mathbf{a}_{(0)})]^2}{N-M}$$

The Bootstrap Method

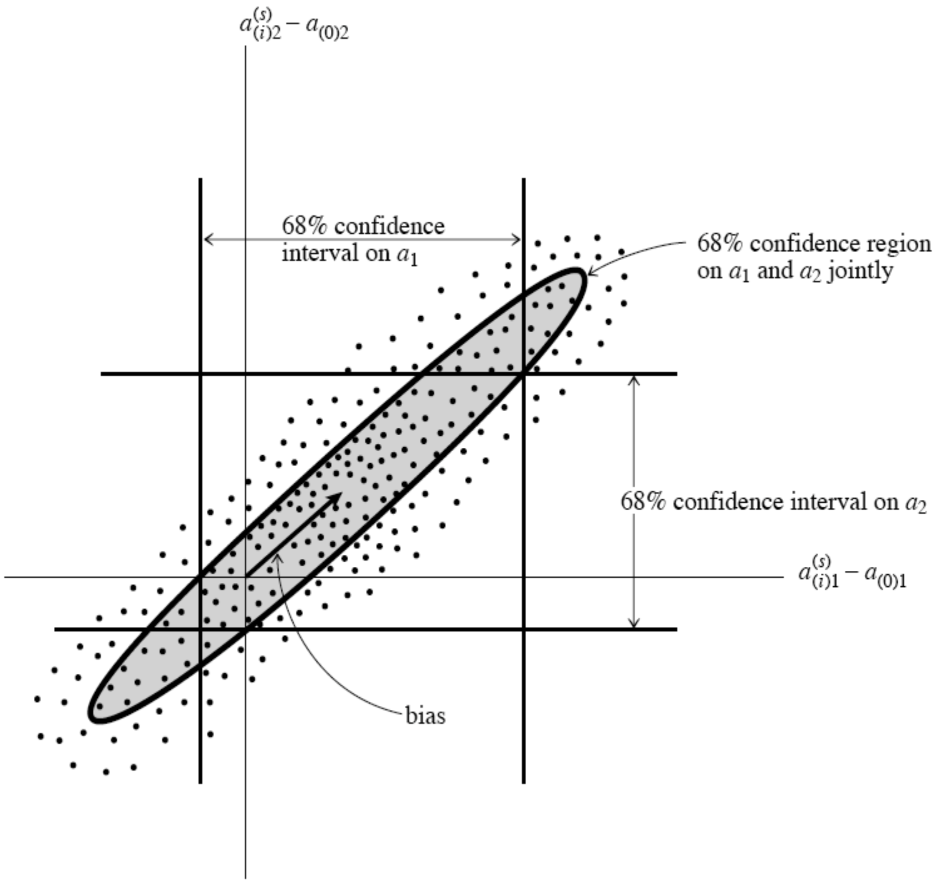
- If you don't know enough about the measurement errors (i.e. cannot even say they are normally distributed) then Monte Carlo simulation cannot be used.

- Bootstrap Method uses actual data set $\mathcal{D}_{(o)}$, with its N data points, to generate synthetic data sets $\mathcal{D}^{S}_{(1)}$, $\mathcal{D}^{S}_{(2)}$,… also with N data points.

- Randomly select N data points from $\mathcal{D}_{(o)}$ *with replacement*, which makes $\mathcal{D}^{S}_{(j)}$ differ from $\mathcal{D}_{(o)}$ with a fraction of the original points replaced by *duplicated* original points.

- The χ^2 merit function does not depend on the order of (x_i,y_i), so fitting the $\mathcal{D}^{S}_{(j)}$ data yields model parameter sets $\mathbf{a}^{S}_{(j)}$ as with Monte Carlo, except using actual measurement noise.

Confidence Intervals and Accuracy of Model Parameters


$a^{(s)}_{(i)2} - a_{(0)2}$
68% confidence interval on a_1
68% confidence region on a_1 and a_2 jointly
68% confidence interval on a_2
$a^{(s)}_{(i)1} - a_{(0)1}$
bias

In MatLab: `y=prctile(x,[5 95])`

- The probability distribution is a function defined on M-dimensional space of parameters **a**.
- A *confidence interval* is a region that contains a high percentage of the total distribution relative to model parameters of interest.
- You choose the confidence level (e.g. 68.3%, 90%, etc.) and the region shape.
  - e.g. lines, ellipses, ellipsoids
- You want a region that is compact and reasonably centered on $\mathbf{a}_{(0)}$.

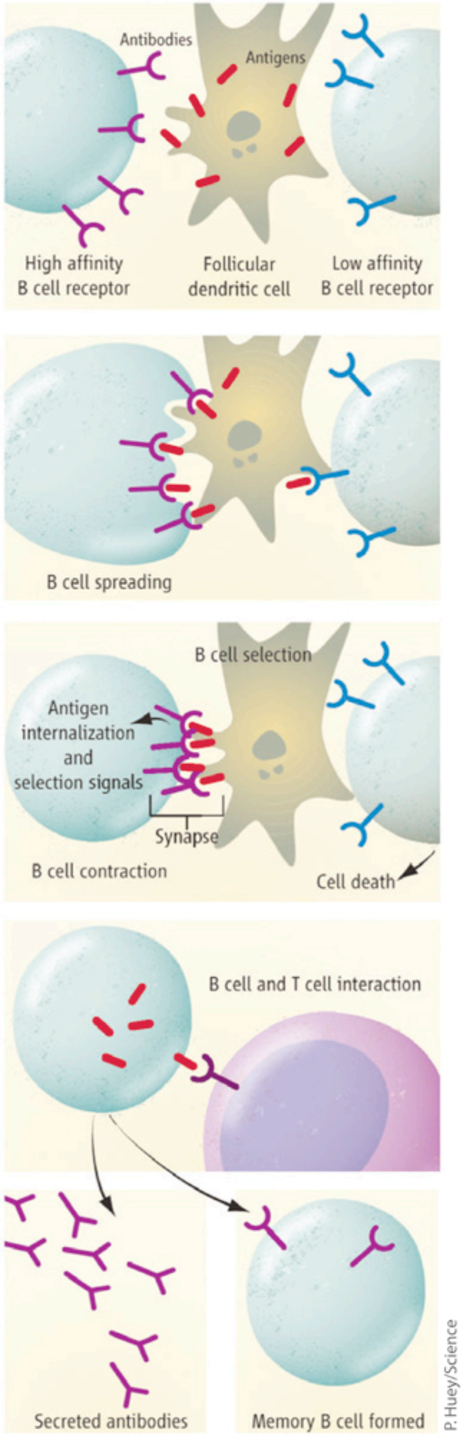
Validating Physical Interpretation of Model Parameters

- Physical sensibility
  - Chemical rate constant cannot be negative
  - Poisson’s ratio cannot exceed 0.5
  - Can enforce lower and upper bounds on parameters, but should examine closely if these end up “optimal”
- Independent measurements of key physical quantities
  - Comparison with published values or limiting behavior
  - Measure steady state modulus of viscoelastic material
- Experimentally alter specific parameters, collect data, and examine results of model fit
  - May involve building a physical model for testing
- Compare model fitting results using data from normal and abnormal populations
  - In asthma patients, airway resistance should be higher than normal

Assignment

B lymphocytes in the immune response

www.EnCognitive.com


Antibodies
Antigens
High affinity B cell receptor
Follicular dendritic cell
Low affinity B cell receptor
B cell spreading
B cell selection
Antigen internalization and selection signals
Synapse
B cell contraction
Cell death
B cell and T cell interaction
Secreted antibodies
Memory B cell formed
P. Huey/Science

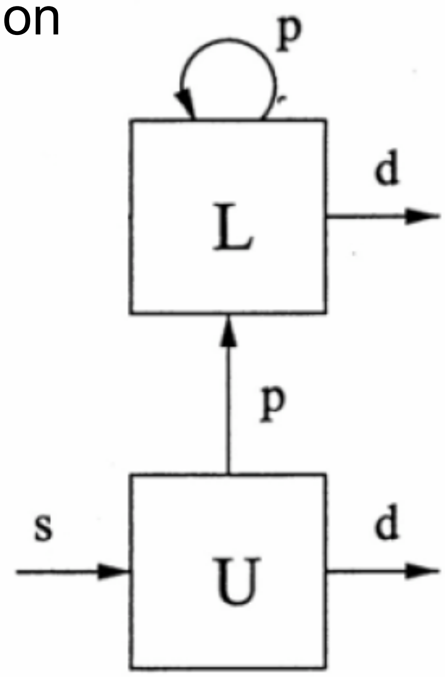
DeBroe, *Kidney Int*, 2006

Assignment

- ODE model of BrdU labeling to estimate proliferation and death rates of B cells.

U – number of unlabeled B cells
L – number of BrdU labeled B cells

p – rate of proliferation (per hour)
d – rate of death (per hour)
s – rate of cell inflow from source (cells/hr)



- Given experimental data on fraction of total B cells labeled with BrdU versus time, develop a model to fit the data, estimate values of *p*, *s*, and *d*, and evaluate the model performance.

Steven Kleinstein and Uri Hershberg

Resources

- Numerical Recipes online

  www.nr.com/nronline_switcher.html

- Matlab online help

  www.mathworks.com/access/helpdesk/help/techdoc/

- References

Anderson SM, Khalil A, Uduman M, Hershberg U, Louzoun Y, Haberman AM, Kleinstein SH, Shlomchik MJ. Taking advantage: high-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells, *J Immunol*, 183:7314-7325, 2009.

Glantz SA. *Primer of Biostatistics*, 6th Ed., McGraw-Hill, 2005.

Lobemeier ML. Linearization plots: time for progress in regression, *BioMedNet*, issue 73, March 3, 2000.

Lutchen KL and Costa KD. Physiological interpretations based on lumped element models fit to respiratory impedance data: use of forward-inverse modeling, *IEEE Trans Biomed Eng*, 37:1076-1086, 1990.