

## 1) Some key concepts:

Physical experiment - measure the relation between a number of experimental *variables*..

Variables may be:

- controlled by us (common in Chemistry)
- not-controlled by us, but selected by us (common in Astronomy and Geology)
- determined by physical laws from other variables (*dependent variables*).

Example: Certain amount of gas in closed cylinder with piston. *Independent variables* = V and T. *Dependent variable* = P. The choice of dependent variable depends on the experimental setup.

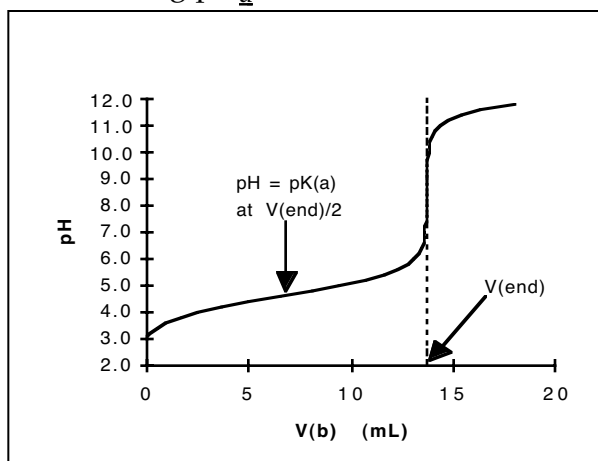
Functional Form: for example  $P = c T/V$ , where c is a constant *parameter*. A proposed functional form may be derived from what is believed to be fundamental physical or chemical laws, or it may be empirical, or somewhere in between. It is customary to write the dependent variable as a function of the independent variables and parameters.

Least Squares Estimation of Parameters - a method of estimating parameters for a functional form which is assumed to be the correct description of the dependence of one (or more) variables on one or more independent variables.

Notation: x denotes independent variable; y denotes dependent variable.  
 $\theta$  denotes a generalized parameter. f denotes generalized function.  
 So ...  $y = f(x; \theta)$ . (There may be several different x ( $x_1, x_2$ , etc.) or  $\theta$ .)

## 2) Some examples of least squares estimation of parameters

Determining  $pK_a$ 's from titration data:



Determine  $pK_a$ 's for polyprotic weak acid  $H_nL$ .

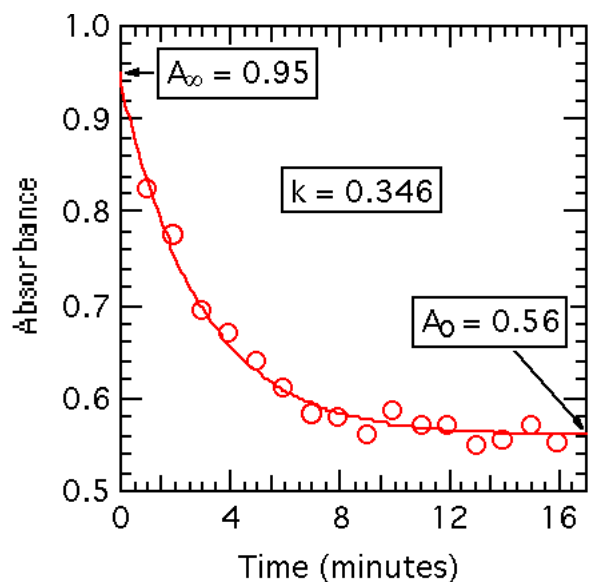
Natural to consider pH to be dependent variable.  $V_{base}$  is independent variable.

$pH = f(V_{base}; V_{initial}, n_L, n_{H,orig}, pK_{a1}, pK_{a2}, \dots, pK_{an})$ .

Sometimes some parameters are known pretty well, but you may want to change them later (such as  $n_L$  and  $n_{H,orig}$ ).

Note that it is easier to calculate  $V_{base}$  from pH rather than *vice versa*. So some people consider  $V_{base}$  as dependent variable; pH is independent. (cf. Chem 100/101 lab)

Determining rate constants for exponential decay



Say the concentration of a species is monitored by observing the absorbance of light at a certain wavelength (remember Beer's Law). Further, assume one colored species reacts to form a second colored species in a first-order reaction (exponential decay).

$$Abs = A_{\infty} + (A_0 - A_{\infty}) e^{-kt}$$

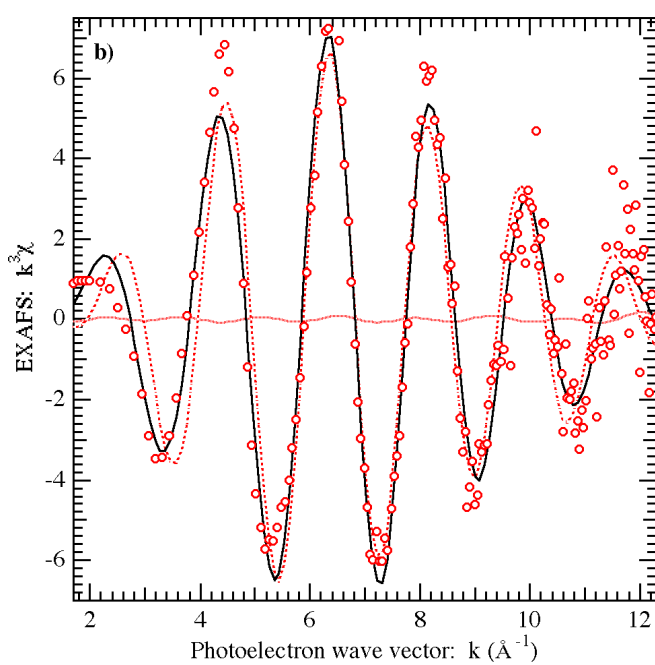
Independent variable = t

Dependent variable = Abs

Parameters =  $A_0, A_{\infty}, k$ (rate constant)

$$[Abs = f(t; A_0, A_{\infty}, k)]$$

Determining bond distances from EXAFS spectroscopy:



$$\chi = \frac{n A(k)}{k r^2 \exp(2\sigma_d^2 k^2)} \sin[2k r + \alpha(k)]$$

The independent variable is k. The dependent variable is  $\chi$  (EXAFS).

$$[\chi = f(k; n, r, \sigma_d)]$$

The parameters are n, r and  $\sigma_d$  (number of atoms around the X-ray absorbing atom, average bondlength and the vibrational and static disorder). (Here  $\sigma$  is *not* a statistical e.s.d.)

A(k) and  $\alpha(k)$  are tabulated functions (for any value of k, can look up).

Start with data set: ca. 200 pairs (k,  $\chi$ ).

The goal is to determine n, r and  $\sigma_d$  as those parameters that give the "best" fit to data. (In graph, solid fit looks better than dashed fit, but how do we quantitate this?)

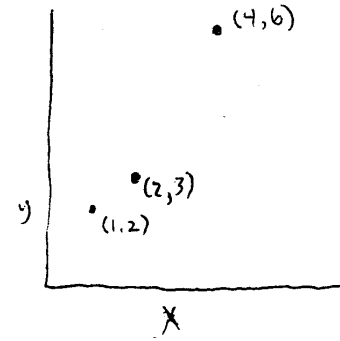
### 3) Start with a simple linear least-squares problem:

What is the best line fit to the data shown; i.e. what are the best choices of  $\theta_1$  and  $\theta_2$  for  $y = \theta_1 + \theta_2 x$ ?

We need a definition of "best".

**Least-squares criterion: The best fit is the fit that minimizes the residual R, where**

$$R = \sum_{\text{data points}} (y_{\text{calc}} - y_{\text{obs}})^2 \quad (\text{In this equation, } y_{\text{calc}} = \theta_1 + \theta_2 x_{\text{obs}}.)$$



Note that the residual  $R$  can be expressed as a function of  $\theta_1$  and  $\theta_2$ :

$$\begin{aligned} R &= (y_{\text{calc}} - y_{\text{obs}})_{x=1}^2 + (y_{\text{calc}} - y_{\text{obs}})_{x=2}^2 + (y_{\text{calc}} - y_{\text{obs}})_{x=4}^2 \\ &= (\theta_1 + \theta_2 - 2)^2 + (\theta_1 + 2\theta_2 - 3)^2 + (\theta_1 + 4\theta_2 - 6)^2 \\ &= \theta_1^2 + \theta_2^2 + 4 + 2\theta_1\theta_2 - 4\theta_1 - 4\theta_2 + \theta_1^2 + 4\theta_2^2 + 9 + 4\theta_1\theta_2 - 6\theta_1 - 12\theta_2 \\ &\quad + \theta_1^2 + 16\theta_2^2 + 36 + 8\theta_1\theta_2 - 12\theta_1 - 48\theta_2 \\ &= 3\theta_1^2 + 21\theta_2^2 + 49 + 14\theta_1\theta_2 - 22\theta_1 - 64\theta_2 \end{aligned}$$

Now, use calculus to find the minimum of  $R$ . Calculus tells us that at such a minimum, the partial derivatives of  $R$  with respect to the parameters must all be zero. For a simple linear least-squares problem, there will always be only one set of parameters where the partial derivatives are zero, and we can use calculus (and algebra) to find that solution:

$$\begin{aligned} \left. \frac{\partial R}{\partial \theta_1} \right|_{\theta_2} = 0 &= 6\theta_1 + 14\theta_2 - 22 & \left. \begin{aligned} &\theta_1 = \frac{11}{3} - \frac{7}{3}\theta_2 \\ &0 = 42\theta_2 + \frac{154}{3} - \frac{98}{3}\theta_2 - 64 \\ &\textcircled{\times 3} \rightarrow \\ &= 126\theta_2 + 154 - 98\theta_2 - 192 \\ &= 28\theta_2 - 38 \\ &\theta_2 = \frac{38}{28} = \frac{19}{14} \Rightarrow \theta_1 = \frac{11}{3} - \frac{7}{3} \cdot \frac{19}{14} \\ &= \frac{22}{6} - \frac{19 \cdot 7}{6} \\ &= \frac{1}{2} \end{aligned} \right\} \\ \left. \frac{\partial R}{\partial \theta_2} \right|_{\theta_1} = 0 &= 42\theta_2 + 14\theta_1 - 64 \end{aligned}$$

Thus  $y = (1/2) + (19/14)x$  is the best fit to the line. The values of  $y_{\text{calc}}$  are shown below:

x	$y_{\text{obs}}$	$y_{\text{calc}}$	$y_{\text{calc}} - y_{\text{obs}}$
1	2	13/7	-1/7
2	3	45/14	3/14
4	6	83/14	-1/14

The minimized residual is

$$R = \left(-\frac{1}{7}\right)^2 + \left(\frac{3}{14}\right)^2 + \left(-\frac{1}{14}\right)^2 = \frac{1}{14}$$

#### 4) A matrix formulation of what we just did, to arrive at a general linear least squares approach.

For linear least squares, the observed  $y$  values ( $y_i$ , where the  $i$  subscript is a counter for all the different measurements of the data set) should conform to equation 1. The  $\varepsilon_i$  represents random measurement error and the other terms are based on a theory.

$$y_i = X_{i1} \theta_1 + X_{i2} \theta_2 + X_{i3} \theta_3 + \dots + X_{ip} \theta_p + \varepsilon_i \quad (\text{eq. 1})$$

The  $X_{ij}$  are either independent variables

functions of independent variables ( $\sin(t)$ ,  $10^{\text{pH}}$ , PV, etc.)

just numbers

**The key requirement is that we know or can calculate in advance the values of  $X_{ij}$ .**

In the graphical example, there are only two parameters ( $p = 2$ ) and

$$X_{i1} = 1 \quad X_{i2} = x_i \quad \theta_1 = \text{x-intercept} \quad \theta_2 = \text{slope}$$

If there are, say, 100 data points, there will be 100 equations like equation 1, each for a different value of  $i$ . That gets hard to keep track of, especially when we are talking in abstract terms. Fortunately, we can write a very simple equation using a matrix formulation.

Here is what you need to know about matrices (you can learn more about them in linear algebra courses).

*Matrices:* An " $r \times c$  matrix" refers to a table of numbers (a spreadsheet, if you wish) with  $r$  rows and  $c$  columns. Matrices are denoted in bold face and/or with double underline: **M**.

*Vectors:* A "vector" is an  $r \times 1$  matrix (or it could be a  $1 \times c$ , but when possible we'll try to arrange numbers in a vector in a vertical arrangement). Vectors are denoted with a single underline. The set of  $y_i$  values is denoted y.

Matrix multiplication is defined as follows:

Matrix multiplication:

$$\underline{C} = \underline{A} \underline{B} \Rightarrow C_{ij} = \text{dot product (} i^{\text{th}} \text{ row of } A) \cdot (j^{\text{th}} \text{ column of } B)$$

$\uparrow$   $\uparrow$   
 $i^{\text{th}} \text{ row}$   $j^{\text{th}} \text{ column}$

if you've never seen dot products before, use the following formula

$$C_{ij} = \sum A_{ik} B_{kj}$$

where the sum is over the  $k$  columns of  $A$  and  $k$  rows of  $B$ .

(It becomes clear from this definition that matrix multiplication  $\mathbf{AB}$  is only possible if  $\mathbf{A}$  has the same number of columns as  $\mathbf{B}$  has rows.)

Three other matrix concepts we'll need are *transpose*, *unity matrix* and *inverse*.

- The transpose of  $\mathbf{A}$ , denoted  $\mathbf{A}'$ , has rows and columns switched so that  $A'_{ij} = A_{ji}$ .
- Unity matrices  $\underline{\mathbf{1}}$  are square matrices with 1's along the diagonal and 0's elsewhere. These have the property that  $\underline{\mathbf{1}} \mathbf{A} = \mathbf{A} \underline{\mathbf{1}} = \mathbf{A}$ .
- In general, inverse matrices can only be found for square matrices, and are denoted  $\mathbf{A}^{-1}$ . The inverse is defined such that

$$\underline{\underline{M}}^{-1} \underline{\underline{M}} = \underline{\underline{1}} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

OK – you are now an expert at matrices. We have a series of  $n$  equations like this:

$$y_i = X_{i1} \theta_1 + X_{i2} \theta_2 + X_{i3} \theta_3 + \dots + X_{ip} \theta_p + \varepsilon_i \quad (\text{eq. 1})$$

This can be rewritten as a matrix equation:

(From the earlier example:

$$\begin{matrix} \boxed{y} \\ \hline \end{matrix} = \begin{matrix} \boxed{X} \\ \hline \end{matrix} \begin{matrix} \boxed{\theta} \\ \hline \end{matrix} + \begin{matrix} \boxed{\varepsilon} \\ \hline \end{matrix} \quad (\text{eq. 2})$$

$n \times 1$        $n \times p$     $p \times 1$

$$\begin{pmatrix} 2 \\ 3 \\ b \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

The residual to minimize is  $R = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 = (\epsilon_1 \ \epsilon_2 \ \epsilon_3) \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix} = \underline{\underline{\epsilon}}' \underline{\underline{\epsilon}}$

↖ transpose  
 ↓ of  $\underline{\underline{\epsilon}}$

$\underbrace{\hspace{10em}}_{1 \times 3 \quad 3 \times 1}$

$$R = \underline{\underline{\epsilon}}' \underline{\underline{\epsilon}} = (\underline{y} - \underline{X}\underline{\theta})' (\underline{y} - \underline{X}\underline{\theta}) = (\underline{y}' - \underline{\theta}' \underline{X}') (\underline{y} - \underline{X}\underline{\theta})$$

the transpose of a matrix product  
 is the product of the transposes,  
 but in reverse order

(Note that  $\frac{X' \theta'}{p \times n \quad 1 \times p}$  is not even defined unless  $n=1$ )

Continuing with the derivation:

$$R = \underline{y}' \underline{y} - \underline{y}' \underline{X} \underline{\theta} - \underline{\theta}' \underline{X}' \underline{y} + \underline{\theta}' \underline{X}' \underline{X} \underline{\theta}$$

transposes of each  
 other are equal because  
 are  $1 \times 1$

$$= \underline{y}' \underline{y} - 2 \underbrace{\underline{\theta}' \underline{X}' \underline{y}}_{\underline{V}} + \underbrace{\underline{\theta}' \underline{X}' \underline{X} \underline{\theta}}_{\underline{M}}$$

For the simple example:

$$\underline{V} = \underline{X}' \underline{y} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 6 \end{pmatrix} = \begin{pmatrix} 11 \\ 32 \end{pmatrix} = \begin{pmatrix} \sum y \\ \sum xy \end{pmatrix}$$

$$\underline{M} = \underline{X}' \underline{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 7 \\ 7 & 21 \end{pmatrix} = \begin{pmatrix} n & \sum x \\ \sum x & \sum x^2 \end{pmatrix}$$

(Note that  $\mathbf{M} = \mathbf{M}'$ .)

Now set the partial derivatives to zero and solve:

$$\frac{dR}{d\theta} = \begin{pmatrix} \frac{\partial R}{\partial \theta_1} \\ \frac{\partial R}{\partial \theta_2} \end{pmatrix} = -2 \underline{X}' \underline{y} + 2 \underline{X}' \underline{X} \underline{\theta} = 0$$

$$-2 \underline{V} + 2 \underline{M} \underline{\theta} = 0$$

$$\underline{M} \underline{\theta} = \underline{V}$$

In order to solve for  $\theta$  (the parameters we are trying to estimate), we left-multiply both sides of the above equation by  $\underline{M}^{-1}$  (finding the inverse of a matrix by hand can be time-consuming, but EXCEL and many other programs contain built-in algorithms for inverting matrices, so we'll just assume that this can be done easily).

$$\underline{M}^{-1} \underline{M} \underline{\theta} = \underline{M}^{-1} \underline{V}$$

$$\underline{\theta} = \underline{M}^{-1} \underline{V} = (\underline{X}' \underline{X})^{-1} \underline{X}' \underline{y}$$

### 5) Now for weighted linear least squares fitting

This is a slight variation of what we just did. Weighting means that we take account of the precision of each measurement of  $y_i$  by estimating a standard deviation of the measurement  $\sigma_i$ . These estimates of  $\sigma_i$  may be all the same, or they may be different for the different data points.

In weighted least squares, the residual is defined as

$$R = \left( \frac{\epsilon_1}{\sigma_1} \right)^2 + \left( \frac{\epsilon_2}{\sigma_2} \right)^2 + \dots + \left( \frac{\epsilon_n}{\sigma_n} \right)^2$$

The weights of each point,  $w_i$ , are defined as  $w_i = \left( \frac{1}{\sigma_i} \right)^2$ .

The weighting matrix,  $\underline{\underline{W}}$ , is a diagonal matrix  $\begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \dots & \\ 0 & & & w_n \end{pmatrix}$

(a *diagonal matrix* is one in which all non-diagonal elements  $[W_{ij}, i \neq j]$  are zero)

A matrix formula for R is  $R = \underline{\underline{\epsilon}}' \underline{\underline{W}} \underline{\underline{\epsilon}}$

$$\underline{\underline{\epsilon}} = \underline{\underline{y}} - \underline{\underline{X}} \underline{\underline{\theta}}, \text{ so } \dots$$

$$R = (\underline{\underline{y}}' - \underline{\underline{\theta}}' \underline{\underline{X}}') \underline{\underline{W}} (\underline{\underline{y}} - \underline{\underline{X}} \underline{\underline{\theta}}) = \underline{\underline{y}}' \underline{\underline{W}} \underline{\underline{y}} - 2 \underline{\underline{\theta}}' \underline{\underline{X}}' \underline{\underline{W}} \underline{\underline{y}} + \underline{\underline{\theta}}' \underline{\underline{X}}' \underline{\underline{W}} \underline{\underline{X}} \underline{\underline{\theta}}$$

Now to find the minimum in R:

$$\frac{dR}{d\theta} = \begin{pmatrix} \frac{dR}{d\theta_1} \\ \frac{dR}{d\theta_2} \\ \vdots \\ \frac{dR}{d\theta_n} \end{pmatrix} = -2 \underbrace{\underline{\underline{X}}' \underline{\underline{W}} \underline{\underline{y}}}_{\text{define } \underline{\underline{V}} = \underline{\underline{X}}' \underline{\underline{W}} \underline{\underline{y}}} + 2 \underbrace{\underline{\underline{X}}' \underline{\underline{W}} \underline{\underline{X}}}_{\text{define } \underline{\underline{M}} = \underline{\underline{X}}' \underline{\underline{W}} \underline{\underline{X}}} = 0$$

$$\Rightarrow \underline{\underline{M}} \underline{\underline{\theta}} = \underline{\underline{V}} \quad \text{or} \quad \boxed{\underline{\underline{\theta}} = \underline{\underline{M}}^{-1} \underline{\underline{V}}}$$

## 6) Estimating the uncertainty of the estimated parameters

One of the reasons least-squares analysis is popular for data analysis is that it provides a method for estimating the uncertainty of each of the estimated  $\theta_i$  values.

$$\text{Variance-covariance matrix } \underline{\underline{S}} = \left( \frac{R}{n_{\text{data}} - n_{\theta}} \right) \underline{\underline{M}}^{-1}$$

where  $S_{ii}$  (diagonal element) =  $(\sigma_i)^2$  ( $\sigma_i$  is e.s.d. of  $\theta_i$ )

correlation coefficient  $c_{ij} = \frac{S_{ij}}{\sigma_i \sigma_j}$  (range between -1 and 1)

*Notes on these matrices:* The diagonal elements of  $\underline{\underline{S}}$  are called variances, while the off-diagonal elements are called covariances of pairs of parameters. A non-zero covariance will give rise to a non-zero correlation coefficient, indicating that the parameter estimates are not independent. A positive correlation coefficient (or covariance) means that if  $\theta_i$  has been estimated too high, then it is likely that  $\theta_j$  has also been estimated too high; a negative correlation coefficient means high estimations of  $\theta_i$  are likely associated with low estimations of  $\theta_j$ . Relatively large correlation coefficients ( $|c_{ij}| > 0.7$ , for instance) are worth noting since they may give insight into the proper interpretation of the results. When the correlation coefficients are near the limits ( $|c_{ij}| > 0.95$ , for instance) the least-square refinement may be less reliable due to round-off errors in the matrix inversion routine.

Often, instead of reporting the minimized residual, researchers report a “goodness of fit” statistic defined as follows:  $GOF = \sqrt{\frac{R}{n_{data} - n_{\theta}}}$ . The GOF should be  $\approx 1$  if the measurement errors (and hence  $\underline{\underline{w}}$ ) have been estimated correctly.

## 7) General (non-linear) least squares refinements

In many cases, the functional form (usually based on some theory or other) for calculating  $y_{calc}$  is not of the form given in equation equation 1. For general least-squares refinements, all that is needed is that there be some function that can be calculated by a computer. It is also helpful if we know enough Calculus to calculate formulae for the partial derivatives, although it is usually easy enough to have the computer estimate these numerically (as  $\Delta f / \Delta \theta_i$  for very small increments of  $\theta_i$ , keeping all other  $\theta$  the same).

$$y_i = f(x_i; \theta_1, \theta_2, \dots, \theta_n) + \epsilon_i$$

In general least squares, one needs to start with guesses for the  $\theta_i$ . Often the algorithm presented here won't work unless the guesses are “good” in that they are based on preliminary analysis of the data (i.e. random guesses won't work). Call the vector of initial guesses  $\underline{\theta}_g$ , and let  $\underline{y}_g$  denote the values of  $y_{calc}$  based on these guesses:

$$y_{g,i} = f(x_i; \theta_{g1}, \theta_{g2}, \dots, \theta_{gn})$$

Assume that  $\underline{\theta}_g$  is close to the “true”  $\underline{\theta}$  and estimate  $\underline{y}$  for these “true” parameters based on the  $\underline{y}_g$  and the partial derivatives  $\delta f / \delta \theta_i$ :

$$f(x_i; \theta_1, \theta_2, \dots, \theta_n) \approx f(x_i; \theta_{g1}, \theta_{g2}, \dots, \theta_{gn}) + \sum_j \left( \frac{\partial f}{\partial \theta_j} \right) (\theta_i - \theta_{gj})$$

↑  
evaluated at  $\underline{\theta}_g$  and  $x_i$

Putting the previous equation in matrix notation:

$$\underline{y} - \underline{\epsilon} \approx \underline{y}_g + \left( \frac{df}{d\theta} \right) (\underline{\theta} - \underline{\theta}_g)$$

$\uparrow$   
 an  $n_x \times n_\theta$   
 matrix  $\Rightarrow$  call it  $\underline{X}$

$$\underbrace{(\underline{y} - \underline{y}_g)}_{\underline{\Delta y}} \approx \underline{X} \underbrace{(\underline{\theta} - \underline{\theta}_g)}_{\underline{\Delta \theta}} + \underline{\epsilon}$$

$$\underline{\Delta y} \approx \underline{X} (\underline{\Delta \theta}) + \underline{\epsilon} \tag{eq. 3}$$

The equation in the last line involves definitions:  $\underline{\Delta \theta} = \underline{\theta} - \underline{\theta}_g$  and  $\underline{\Delta y} = \underline{y} - \underline{y}_g$ . These vectors represent the shifts in parameter estimates and  $y_{\text{calc}}$  values that need to be made. General least squares is based on the observation that equation 3 has the same form as equation 2 from the linear least squares derivation. There are some differences:  $\underline{y}$  has become  $\underline{\Delta y}$ ,  $\underline{\theta}$  has become  $\underline{\Delta \theta}$ , and the equation is only approximate. But for now, use the methods of the weighted linear least-squares section to solve for  $\underline{\Delta \theta}$ :

$$\underline{\Delta \theta} = (\underline{X}' \underline{W} \underline{X})^{-1} \underline{X}' \underline{W} \underline{\Delta y} \tag{eq. 4}$$

After solving for  $\underline{\Delta \theta}$ , calculate new guesses for  $\underline{\theta}$  as  $\underline{\theta}_{\text{new}} = \underline{\theta}_g + \underline{\Delta \theta}$ .

These new guesses will probably still not be the best guesses (they won't minimize R) because of the approximation in equation 3. But usually they are closer. So with the new guesses, calculate a new set of  $\underline{y}_g$  and  $\underline{X}$  (based on partial derivatives), and then use equation 4 to calculate a new shift in parameters. Each go-around is called a "cycle" of least squares refinement.

After each cycle, one can calculate statistics on the estimated parameters; for instance  $\text{esd}(\underline{\theta}) = \text{esd}(\underline{\Delta \theta})$ . The refinement is continued until the shift in each parameter is less than 10% of the esd of that parameter. (Some programs stop when the fractional decrease in R from one cycle to the next is less than a certain cut-off value).

That's it for the theory of least squares fitting. Like most theories, it is hard to understand until you implement it in practice. For this reason, I have developed a companion handout giving directions for using EXCEL to perform linear and general least-squares calculations. Following the directions in this handout will help you to understand what matrix multiplication looks like in practice.