The Theory of Least Squares Determination of Variables 4/1/05

Kenneth N. Raymond
University of California, Department of Chemistry
Berkeley, California

Glossary of Symbols

[Note that subscripts denote individual quantities (a number, data point), vectors will be written as bold faced lower case letters (e.g., **v**) and matrices as bold faced upper case (e.g. **M**)].

y_i^o Observed quantity, experimental data.

y_i^c The corresponding calculated quantity from a theoretical model.

p_j A variable parameter of y^c which is to be adjusted to get y to agree with y^o

x_j A variable of y^c that is not a parameter to be fit.

n Number of observations.

m Number of parameters to be fit.

Introduction

The following notes present the basic equations for least squares procedures. The notation is primarily in matrix form since the equations are most clearly and concisely expressed in this way.

Several examples are given to provide familiarity with the actual process of least squares. For further discussion and more complete proofs in this area the book by Walter Hamilton is especially recommended.

I. LINEAR EQUATIONS

Suppose there are n observations and also n parameters where x_{ij} is the j^{th} parameter for the i^{th} observation and we know that y_i is linearly related to all the x_j's so that

1) $$y_i = p_1 x_{i1} + p_2 x_{i2} + \ldots + p_n x_{in}$$

or in matrix form: **y = Xp** (p_j are coefficients, or variable parameter to be fit) **X** = {x_{ij}}).

Orders of matrices: n x 1, n x n, and n x 1 respectively.

Then if the determinant of $\mathbf{X} \neq 0$, $\mathbf{X}^{-1}$ exists such that $\mathbf{X}^{-1}\mathbf{X} = \mathbf{X}\mathbf{X}^{-1} = \mathbf{I}$ (identity matrix), and thus:

2) $$\mathbf{X}^{-1}\mathbf{y} = \mathbf{X}^{-1}\mathbf{X}\mathbf{p} = \mathbf{I}\mathbf{p} = \mathbf{p}$$

or $\mathbf{p} = \mathbf{X}^{-1}\mathbf{y}$. So we have solved for the coefficients, p_j, that related the variables x_{ij} to the y_i observations.

II. LINEAR LEAST SQUARES

Determination of Variables Which Give "Best Fit"

Suppose there are now n observations and m parameters. (If $n < m$ the problem can't be solved. If $n = m$ the problem is that just described.)

We now have

3) $$y_i^c = \sum_{j=1}^{m} p_j x_{ij} = \sum_{j=i}^{m} x_{ij} p_j$$

or	$\mathbf{y}^c$	=	$\mathbf{X}$	•	$\mathbf{p}$
	n x 1		n x m		m x 1
	variables				coefficients (i.e., fit parameters) to be determined

Now these y_i^c are **calculated** values. Let's call the corresponding observations y_i^o. Then we want to minimize the disagreement between y_i^c and y_i^o.

Define a residual, R, as:

4) $$R = \sum_{i=1}^{n} (y_i^c - y_i^o)^2 \text{ and we want}$$

5) $$\frac{\partial R}{\partial p_j} = 0 = 2\sum_{i=1}^{n} \left(y_i^{\mathrm{c}} - y_i^{\mathrm{o}}\right)\frac{\partial y_i^{\mathrm{c}}}{\partial p_j}$$

We will state without proof that if the variance of each y_i^{o} is the same, the variances of the calculated parameters, p_i, are minimized when the sum of the differences between y_i^o and y_i^c **squared** is minimum. Contrary to a popular misconception, **this result is independent of the probability distributions of the errors in** y_i^{o}. Furthermore, the variance of any derived quantity which is a linear combination of the calculated parameters, p, is a minimum when R is a minimum. This is what we mean when we say that the **best estimate** of the parameters is obtained by a least squares fit. (For a discussion of this point see any book on statistics. Walter Hamilton's book, *Statistics in Physical Science*, Ronald Press, 1964, is again especially recommended.)

Let us now write R as:

6) $$R = (\mathbf{Xp} - \mathbf{y}^{\mathrm{o}})^{\mathrm{T}} (\mathbf{Xp} - \mathbf{y}^{\mathrm{o}})$$ (Note that for matrices **A** and **B**, $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}}$)

$$\mathrm{R} = \mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Xp} - 2\mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}^{\mathrm{o}} + \mathbf{y}^{\mathrm{o}^{\mathrm{T}}}\mathbf{y}^{\mathrm{o}}$$

For optimum p_j's we have:

7) $$\frac{\partial \mathrm{R}}{\partial p_1} = 0\,,\ \frac{\partial \mathrm{R}}{\partial p_2} = 0\,,\ \ldots\ \frac{\partial \mathrm{R}}{\partial p_n} = 0$$

or $$\frac{\partial \mathrm{R}}{\partial \mathbf{p}} = \begin{pmatrix} \frac{\partial R}{\partial p_1} \\ \frac{\partial R}{\partial p_2} \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \end{pmatrix}$$

m x 1 m x 1

We may differentiate equation 4 in its summation form or (equivalently) differentiate the vector and matrix equation in equation 6 to give:

8) $$\frac{d\mathrm{R}}{d\mathbf{p}} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y}^{\mathrm{o}} = 0$$

or

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p} = \mathbf{X}^{\mathrm{T}}\mathbf{y}^{\mathrm{o}}$$

for the optimum set of parameters, **p**.

[You may check this shorthand method of differentiation by actually writing out the partial derivative expressions for the p_j's from equation 4. You will then find that you have a set of equations which correspond to what we have just derived. This is recommended as an exercise.]

The product $\mathbf{X}^{\mathrm{T}}\,\mathbf{X}$ is a symmetric m x m matrix (why?), and if $|\mathbf{X}^{\mathrm{T}}\,\mathbf{X}| \neq 0$ this matrix has an inverse, so that we may write:

$$\mathbf{p} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathrm{T}}\mathbf{y}^{\mathrm{o}}$$

To repeat, this means that is we have a postulated functional relationship

$$y_i^c = p_1 x_{i1} + p_2 x_{i2} + \ldots + p_n x_{im}$$

and we observe

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \cong \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & & \\ \vdots & & & & \\ x_{n1} & \cdots & \cdots & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix}$$

$\mathbf{y}^{\mathrm{o}}$	$\mathbf{X}$	$\mathbf{p}$
observed quantities	matrix of variables	unknown coefficients

The optimum set of coefficients, p_j, as calculated from the observations, y_i, found at values of the variables, x_{ij}, is just given by equation we have written.

Estimation of Errors

Now suppose we ask what error limit we may assign to our calculated values on the basis of their agreement with the observations. Define the **variance** of a variable as the average (mean) value of the square of the deviation of that variable. (This will be the square of the standard deviation.) Define the **covariance** of two variables as the mean of the product of the deviations of the two variables. (This product will average to zero if the variables are independent.) We state without proof that the variance-covariance matrix is given by

10) $$\mathbf{S} = \left(\frac{\mathrm{R}}{\mathrm{n-m}}\right)(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$$

that is, it is just related to the inverse matrix by a scale factor (which is the agreement factor divided by the number of observations minus the number of variance parameters). The elements of the variance-covariance matrix give the average of the product of the deviation for the two variables that correspond to the row and column of each element in the matrix. Thus the diagonal elements (variable *i* with variable *i*) give the variance and the off-diagonal elements (variable *i* with variable *j* [$\neq$ *i*]) give the covariance. The elements of this matrix are:

$$\mathbf{S} = \{s_{ij}\} \qquad \mathbf{s}_{ij} = \sigma_i \sigma_j c_{ij} = \text{covariance (or variance)}$$

where σ_i is the **standard deviation** of the i^{th} parameter and c_{ij} is the **correlation coefficient** between the variables numbered *i* and *j*. Clearly $c_{ii} = 1$ since a parameter is directly correlated with itself. The diagonal terms of **S** are just the variances of the parameters:

$$s_{ii} = \sigma_i^2 = \sigma^2(p_i)$$

The **correlation matrix** is the matrix of c_{ij}'s

11) $$\mathbf{C} = \{c_{ij}\} \qquad \mathrm{c_{ij}} = \frac{s_{ij}}{\sigma_i \sigma_j}$$

Example 1

Suppose we have solutions of two compounds that interact such that we always have a mixture. They both absorb at a given wavelength but we can never get a solution of just one compound to determine its extinction coefficient at this wavelength. From Beer's law we show that the measured absorbance is the sum of the absorbance for the two compounds, i.e.

$$\text{Abs} = c_1\, \varepsilon_1 + c_2\, \varepsilon_2$$

where the concentrations of compounds 1 and 2 are c_1 and c_2 and the extinction coefficients are the corresponding ε values. If we recast this in the notation we have used for least squares the extinction coefficients become our parameters (p_1, p_2) and the experimental variables (concentrations) form the matrix **X** The absorbance values are the $\mathbf{y}^o$ values. To solve for the best estimate of the extinction coefficients we get the product matrix ($\mathbf{X}^T\mathbf{X}$), invert this, and multiply by ($\mathbf{X}^T\mathbf{y}^o$) to get **p**, the best least squares fit for the extinction coefficients.

X (matrix of concentrations)		$\mathbf{y}^{obs}$	$\mathbf{y}^c$
0.050	0.100	0.842	.869
0.050	0.050	.480	.472
0.050	0.015	.197	.195
0.050	0.025	.295	.274
0.100	0.100	.968	.945
0.100	0.050	.547	.549
0.100	0.015	.261	.271
0.100	0.025	.338	.351

We then get the values

$(\mathbf{X}^T\mathbf{X})$		$(\mathbf{X}^T\mathbf{X})^{-1}$		$\mathbf{X}^T\mathbf{y}^o$	**p**	σ
0.050	0.029	51.08	-54.52	.3021	1.52	.13
0.029	0.027	-54.52	94.65	.2550	7.93	.18

Remembering that $\mathbf{p} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}^o$ we solve for **p**. We then calculate the values of $\mathbf{y}^c$ from $\mathbf{y}^c = \mathbf{X}, \mathbf{p}$. From the squared differences of $\mathbf{y}^o$ and $\mathbf{y}^c$ we compute a residual, R =

0.018. This times the inverse matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ gives the variance-covariance matrix. The diagonal elements give the standard deviations.

The correlation between the extinction coefficients is -.78. [Why should we expect rather high correlation? Why is it negative?]

Propagation of Error

The only reason one is normally concerned with correlation coefficients is in the correct propagation of errors for quantities derived from parameters resulting from a least squares analysis. For linear functions the propagation of error is very simple.

Example 2

For example, if a derived quantity b is related to the parameters p_1 and p_2 by $b = q_1p_1 + q_2p_2$ then

$$\sigma^2(b) = q_1{}^2\sigma^2(p_1) + q_2{}^2\sigma^2(p_2) + 2q_1q_2c_{1,2}\sigma(p_1)\sigma(p_2)$$

we can see from this equation that large correlation coefficients strongly affect the derived standard deviations.

In application to X-ray crystallography, we may have performed a least squares analysis, in a manner yet to be discussed, to fit an atom which lies on a symmetry axis so that only one parameter, p, is varied: Let's assume we are measuring p in Å and we find its value is +1.01 (2). Suppose there is a center of symmetry at the origin such that

The distance between the two atoms, which we will call a bond length, is clearly 2.02 Å. Now what is the standard deviation for this bond length? Since atom 2 is related to 1 by the center of symmetry:

$p_{atom2} = - p_{atom1}$

and

$c_{1,2} = -1$ Bond $= p_1 - p_2$, so $q_{-1} = 1$, and $q_2 = -1$

$\sigma^2_{bond} = (1)^2(0.02)^2 + (-1)^2(0.02)^2 + (2)(1)(-1)(-1)(0.02)(0.02)$

$\sigma^2_{bond} = 4(0.02)^2$

$\sigma_{bond} = 2(0.02) = 0.04$ Å

Bond = 2.02(4) Å or twice the absolute error in the position of each atom. Notice how physically reasonable this is, since a shift caused by an error is doubled in its contribution to the bond length.

A generalized form for the variance in a linear function of the parameters obtained from a least squares analysis can be expressed most simply in matrix form. If b is the calculated value for a linear function of the variables, p (obtained in a least squares analysis where the coefficients that linearly related the variable p to the parameter b and q) then:

$$b = \sum_{j=1}^{m} q_j p_j = \mathbf{q}^T \mathbf{p}$$

$$\sigma^2(b) = \sum_j \sum_k q_j q_k s_{jk} = \mathbf{q}^T \mathbf{S} \mathbf{q}$$

Where s_{jk} is the element of the variance-covariance matrix for p.

$$s_{jk} = \sigma(p_j)\sigma(p_k)c_{jk}$$

where c_{jk} is the correlation coefficient between j and k. We may write:

$$\sigma^2(b) = \mathbf{q}^T \mathbf{S} \mathbf{q}$$

We may in general have a non-linear dependence of the calculated parameter, b, on the m variables, **p**. That is

$$b = f(\mathbf{p})$$

Then, for small errors we can write

$$q_j = \frac{\partial b}{\partial p_j}$$

and we still have

12) $\sigma^2(b) = \mathbf{q}^T \mathbf{S} \mathbf{q}$

If we have a whole set of derived parameters, b_j and $b_j = f_j(\mathbf{p})$, then the new variance-covariance matrix for the parameters b_j is given by:

13) $\mathbf{Q}^T \mathbf{S} \mathbf{Q}$

where

14) $Q_{ij} = \dfrac{\partial b_j}{\partial p_i}$

Example 3

A common linear least squares problem is the best line through a series of points. That is, we assume:

$y_i = p_1 + p_2 x_i$

We call p_2 the slope and p_1 the intercept of the line. The linear least squares equations in this case assume a particularly simple form. We have:

$$\begin{pmatrix} y_1^c \\ y_2^c \\ \vdots \\ y_n^c \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$$

n x 1 n x 2 2 x 1

and we know that, for the best fit:

$$p = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}^o$$

This gives us:

$$\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

Note that:

$$\mathrm{Det}(\mathbf{X}^T\mathbf{X}) = n\sum x_i^2 - (\sum x_i)^2$$

Invert the matrix:

$$\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \frac{1}{(n\sum x_i^2 - (\sum x_i)^2)} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

p $(\mathbf{X}^T\mathbf{X})^{-1}$ $\mathbf{X}^T\mathbf{y}^o$

The variance-covariance matrix, S, is then

$$S=\frac{R}{(n-2)(n\sum x_i^2-(\sum x_i)^2)}\begin{pmatrix}\sum x_i^2 & -\sum x_i \\ -\sum x_i & n\end{pmatrix}$$

where $R=\sum (y_i^c-y_i^o)^2=\sum (p_1+p_2x_i-y_i^o)^2=np_1+np_2\sum x_i-\sum x_i^o$

and so

$$s_{11}=\sigma^2(p_1)=\left(\frac{R}{(n-2)(n\sum x_i^2-(\sum x_i)^2)}\right)(\sum x_i^2)$$

gives the variance and standard deviation of the intercept

$$s_{22}=\sigma^2(p_2)=\left(\frac{R}{(n-2)(n\sum x_i^2-(\sum x_i)^2)}\right)(n)$$

gives the variance (or standard deviation) of the slope.

The correlation coefficient between the slope and intercept is:

$$c_{12}=s_{12}/(s_{11}s_{22})^{1/2}=\frac{-\sum x_i}{(n\sum x_i^2)^{1/2}}\text{ -}$$

Some qualitative statements can be made by inspection of these formulas:

a) The correlation coefficient approaches +1 as all observed points converge. That is, the observations should be as widely spaced in x as possible.

b) The determinant increases (the problem becomes better conditioned) when equal numbers of observations are made in negative x regions. Note that then $\Sigma x_i=0$. This lowers the standard deviations and the correlation coefficients.

c) As expected, the standard deviations of the slope and intercept of the least squares line increase with R, the degree of fit, and decrease with n, the number of observations.

III. WEIGHTED LEAST SQUARES

Often some observations are more precise than others. We want to emphasize their contribution to the derived parameters relative to less accurate data. Thus we wish to weigh the least squares fit such that

15) $$R = \sum_i \left(\frac{y_i^c - y_i^o}{\sigma_i} \right)^2$$

where σ_i is that standard deviation assigned to the observation $\mathbf{y_i^o}$.

Then $R = [\mathbf{V}(\mathbf{Xp} - \mathbf{y}^o)]^T [\mathbf{V}(\mathbf{Xp} - \mathbf{y}^o)] = \mathbf{p}^T \mathbf{X}^T \mathbf{V}^T \mathbf{V X p} - 2\mathbf{p}^T \mathbf{X}^T \mathbf{V}^T \mathbf{V y}^o + (\mathbf{V y}^o)^T \mathbf{V y}^o$

where

$$V_{ii} = \frac{1}{\sigma_i} \quad , \quad \mathbf{y}_i^c = \mathbf{Xp}$$

and

$$\mathbf{V_{ij}} = 0 \text{ if } i \neq j$$

$$\frac{dR}{dp} = 2\mathbf{X}^T\mathbf{V}^T\mathbf{VXp} - 2\mathbf{X}^T\mathbf{V}^T\mathbf{Vy}^o$$

and we have

16) $$\mathbf{p} = (\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{Wy}^o$$

where $\mathbf{W} = \mathbf{V}^T\mathbf{V} = \{1/s_i^2\}$ is a diagonal matrix of the assigned weights. This whole equation has the same form as the unweighted refinement (equation 9) if we let

$$\mathbf{y}_{\text{new}}^o \equiv \{y_i^o / \sigma_i\} \quad , \quad \mathbf{y}_{\text{new}}^c \equiv \{y^c / \sigma_i\} \quad , \text{ and } \quad \mathbf{X}_{new} = \{X_{ij} / \sigma_i\} .$$

IV. NON-LINEAR LEAST SQUARES IN X-RAY CRYSTALLOGRAPHY

Suppose that we have trial set of parameters for the crystal structure. We denote these by the vector ($\mathbf{p}^o$) whose elements are just these trial parameters. Our calculated structure factors are clearly functions of these parameters, i.e.

$F_i^{calc} = f(p_i^o)$ for a given (i^{th}) value of h, k, and l.

But in general these are close to, but not the same as, the optimum set of parameters that gives the lowest R. Suppose we call the parameters p those which give the best agreement, that is

17) $$R = \sum_i (|F_i^c| - |F_i^o|)^2$$

is a minimum for the parameter set **p**. Set

18) $$\mathbf{p} = \mathbf{p}^o + \Delta\mathbf{p}$$

Where $\mathbf{p}_o$ is our trial set of parameters. We may do a Taylor's series expansion to get:

19) $$F_i^{calc}(\mathbf{p}) = F_i^{calc}(\mathbf{p}^o) + \sum_{j=1} \frac{\partial F_i}{\partial p_j}\Delta p_j \text{ + higher order terms}$$

Neglect the higher order terms (which contain Δp_j^2, etc. and which for small Δp will go to zero) thereby **linearizing** the equation. This has exactly the same form as the previous linear least-squares problem if

$$y_i \equiv F_i^c - F_i^o \quad , \quad D_{ij} = \left(\frac{\partial F_i}{\partial p_j}\right)_{p^o} \quad \text{with } \mathbf{X} \equiv \mathbf{D}$$

and so we know the solution is

20) $$\Delta\mathbf{p} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T(\mathbf{F}^c - \mathbf{F}^o)$$

Where we are now solving for **changes** in parameters. Remember that F_i^c is the ith calculated structure factor for a given set of **trial** atom parameters and F_i^o is the magnitude of the structure factor as actually observed. In a typical modern structure analysis there may be several thousand observed structure factors. Remember that the elements of the matrix D are simply the derivatives of each structure factor with respect to each parameter. That is:

$$\mathbf{D} \equiv \{D_{ij}\}, \quad D_{ij} = \frac{\partial F_i}{\partial p_j}$$

A typical small molecule structure might have

$\Delta\mathbf{p}$	=	$(\mathbf{D}^T\mathbf{D})^{-1}$	$\mathbf{D}^T$	$(\mathbf{F}^c - \mathbf{F}^o)$
200x1		200x200	200x2000	2000x1

After we calculate the incremental change for each parameter we set

$$\underset{\text{new}}{\mathbf{p}_o} = \underset{\text{old}}{\mathbf{p}_o} + \Delta\mathbf{p}$$

we then plug our new $\mathbf{p}_o$ back into the mill to calculate new derivatives, and so forth. If the refinement converges, as we hope, we eventually come to a point after several cycles of this process where the change in parameters that is calculated is negligibly small. At this point we say the refinement has converged and we stop. This is what is meant by "we performed 8 cycles of least squares at which point the refinement had converged." At this point the linearization **approximation** also becomes an **identity**. The residual or R factor is often taken as an indication of how successful the refinement has been. A common unweighted R factor is

$$21) \qquad R = \frac{\sum \left\| F_i^c \mid - \mid F_i^o \right\|}{\sum \mid F_i^o \mid}$$

R factors between 0.10 and 0.20, or 10 and 20%, usually indicate a correct structure, except for very heavy atom structures in which case one whole atom could be missing at this level. R factors of less than 0.10 or 10% indicate the structure is well refined and, in the absence of some special kind of trouble, is correct.

Our treatment up to this point has assumed all of the errors in our observations were equal. This is clearly not the case. For the counter data the statistical counting error for N counts is $\sqrt{\mathrm{N}}$. The percentage error is then $100/\sqrt{\mathrm{N}}$ and we see that the percentage error for weak reflections is very much greater than for strong ones. We wish to take this into account when doing our least squares fit. That is, we want to weigh the contribution of those observations with small errors more heavily than those with large errors. If σ_i is the estimated standard deviation of F_i^o then we may do the statistically correct thing by minimizing

$$22) \qquad \sum_i \left(\frac{\mid F_i^c \mid - \mid F_i^o \mid}{\sigma_i} \right)^2$$

you will recall that our previous equation for the parameter shifts was:

$$\Delta\mathbf{p} = (\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathrm{T}}(\mathbf{F}^{\mathrm{c}} - \mathbf{F}^{\mathrm{o}})$$

The new equation for weighted least squares is:

23) $$\Delta\mathbf{p} = (\mathbf{D}^{\mathrm{T}}\mathbf{W}\mathbf{D})^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{W}(\mathbf{F}^{\mathrm{calc}} - \mathbf{F}^{\mathrm{obs}})$$

where

$$\mathbf{W} = \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & \cdots \\ 0 & 1/\sigma_2^2 & 0 & \cdots \\ 0 & 0 & 1/\sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

This is the actual equation for least squares analysis. When people speak of the weighting scheme they used, they are describing how they estimated the σ_i's. (Don't confuse these with the errors in the parameters as eventually calculated from the least squares fit.)

The most commonly used R factor which includes the weighting scheme, called the **weighted R factor**, is:

24) $$R_2 = \left\{ \frac{\sum \left[\left(|F_i^{\mathrm{c}}| - |F_i^{\mathrm{o}}| \right) / \sigma_i \right]^2}{\sum \left(F_i^{\mathrm{o}} / \sigma_i \right)^2} \right\}^{1/2}$$

A good measure of the correctness of the standard deviations assigned to the data is the error in an observation of unit weight:

25) $$E = \left\{ [1/(n-m)] \sum \left(\frac{|F_i^{\mathrm{c}}| - |F_i^{\mathrm{o}}|}{\sigma_i} \right)^2 \right\}^{1/2}$$

This number should be 1 if all of the weights are correctly assigned and should be uniform for various classes of the data.

<u>Concepts</u>

Least squares "best fit"

Shifts of parameters

esd's of parameters

Correlation coefficients

Error in observation of unit weight