

Least-Squares (LSQ) fit

Robert Estalella

2011 September

1 Linear least-squares fit

1.1 Observations with different errors ϵ_i

Let us assume that we have n observations y_i^{obs} at points x_i , $i = 1, \dots, n$. The values x_i are supposed to be exact, while the rms error associated with each point observed y_i^{obs} is ϵ_i .

We are searching the model that best fits the observations,

$$y^{\text{mod}}(x) = y^{\text{mod}}(x; a_1, \dots, a_m) = \sum_{k=1}^m f_k(x) a_k,$$

with a linear dependence on m free parameters, a_k , and where $f_k(x)$ are known functions of x . The best fit minimizes the objective function

$$\chi^2 = \sum_{i=1}^n \left[\frac{y_i^{\text{obs}} - y^{\text{mod}}(x_i; a_1, \dots, a_m)}{\epsilon_i} \right]^2.$$

Let us define the Jacobian array J with n (number of data points) rows and m (number of parameters) columns

$$J = \begin{pmatrix} f_1(x_1)/\epsilon_1 & \cdots & f_m(x_1)/\epsilon_1 \\ \vdots & & \vdots \\ f_1(x_n)/\epsilon_n & \cdots & f_m(x_n)/\epsilon_n \end{pmatrix}.$$

(Note that J is a Jacobian, since $f_k(x) = \partial y^{\text{mod}} / \partial a_k$.) The array $J^t J$ is a symmetric $m \times m$ square array, given by

$$J^t J = \begin{pmatrix} \sum_{i=1}^n f_1(x_i) f_1(x_i) / \epsilon_i^2 & \cdots & \sum_{i=1}^n f_1(x_i) f_m(x_i) / \epsilon_i^2 \\ \vdots & & \vdots \\ \sum_{i=1}^n f_1(x_i) f_m(x_i) / \epsilon_i^2 & \cdots & \sum_{i=1}^n f_m(x_i) f_m(x_i) / \epsilon_i^2 \end{pmatrix},$$

and its inverse is called the covariance array $C = (J^t J)^{-1}$. The data column array Y (n rows) is

$$Y = \begin{pmatrix} y_1/\epsilon_1 \\ \vdots \\ y_n/\epsilon_n \end{pmatrix},$$

and the parameter column array A (m rows) is

$$A = (a_k) = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}.$$

The least-squares solution for the parameter array is then given by

$$A = C J^t Y,$$

and the uncertainties in the parameter values are given by the diagonal elements of the covariance array,

$$\epsilon^2(a_k) = C_{kk}.$$

1.2 Observations with the same error ϵ

In the case where all errors are equal, $\epsilon_i = \epsilon$, ($i = 1, \dots, n$) the error ϵ can be taken out of the equations. That is, the Jacobian becomes

$$J = \begin{pmatrix} f_1(x_1) & \cdots & f_m(x_1) \\ \vdots & & \vdots \\ f_1(x_n) & \cdots & f_m(x_n) \end{pmatrix},$$

the data array becomes

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and the solution is, as before,

$$A = CJ^t Y.$$

The error ϵ appears only in the parameter uncertainties,

$$\epsilon^2(a_k) = \epsilon^2 C_{kk}.$$

2 Non-linear least-squares fit

Let us assume that the model $y^{\text{mod}}(x; a_1, \dots, a_m)$ does not depend linearly on the parameters a_k . The simplest way to deal with this case is to linearize the model function,

$$y^{\text{mod}}(x; a_1 + \Delta a_1, \dots, a_m + \Delta a_m) \simeq y^{\text{mod}}(x; a_1, \dots, a_m) + \sum_{k=1}^m \frac{\partial y^{\text{mod}}}{\partial a_k} \Delta a_k,$$

and proceed iteratively. Let us assume that at iteration j the values estimated for the parameters are a_k^j . For the next iteration, the values of the parameters will be $a_k^j + \Delta a_k$, with Δa_k determined from the lsq linear fit

$$y_i^{\text{obs}} - y^{\text{mod}}(x_i; a_1^j, \dots, a_m^j) \simeq \sum_{k=1}^m \frac{\partial y^{\text{mod}}}{\partial a_k} \Delta a_k.$$

The lsq solution for Δa_k is given by

$$\Delta A = CJ^t \Delta Y,$$

where C is the covariance array,

$$C = (J^t J)^{-1},$$

J is the Jacobian,

$$J = \begin{pmatrix} \frac{1}{\epsilon_i} \frac{\partial y^{\text{mod}}(x_i; a_1^j, \dots, a_m^j)}{\partial a_k} \end{pmatrix},$$

and ΔY is the ‘‘observation–model’’ column array,

$$\Delta Y = \begin{pmatrix} \frac{1}{\epsilon_i} [y_i^{\text{obs}} - y^{\text{mod}}(x_i; a_1^j, \dots, a_m^j)] \end{pmatrix}.$$

Once ΔA is found, the new values of the parameters will be $a_k^{j+1} = a_k^j + \Delta a_k$, and we can proceed with the next iteration, recomputing $\partial y^{\text{mod}}/\partial a_k$ and $y_i^{\text{obs}} - y^{\text{mod}}(x_i; a_1^{j+1}, \dots, a_m^{j+1})$.

The process is stopped when the parameter increments Δa_k are small enough. The uncertainty in the parameter values, $\epsilon(a_k)$, can be taken equal to those of their increments in the last iteration, $\epsilon(\Delta a_k)$.

3 Uncertainty in the parameter values

In general, when fitting some parameter-dependent model to a set of observations, it is straightforward to find the values of the parameters for which the rms residual, ϵ_{fit}

$$\epsilon_{\text{fit}}^2 = \frac{1}{n} \sum_{i=1}^n [y_i^{\text{obs}} - y^{\text{mod}}(x_i)]^2$$

is minimum. But the problem is to estimate the uncertainty in the values derived for the parameters when there is no analytical expression for the Jacobian array J .

3.1 Montecarlo

One approach is the Montecarlo method. We add to the observation points y_i^{obs} a Gaussian noise of zero mean and standard deviation equal to their rms error ϵ_i , and calculate for each trial the parameter values. The rms deviation of the parameter values obtained is the rms uncertainty $\epsilon(a_k)$ of each parameter.

3.2 Rms residual

A different approach, less computing demanding, is to calculate the rms residual for different values of the parameters, $a_k + \Delta a_k$, one parameter each time. For each parameter, the value of Δa_k for which the rms residual exceeds the minimum residual by a certain amount can taken as the uncertainty in the parameter, $\epsilon(a_k) = \Delta a_k$. What amount is correct?

Let us examine the analytical case, outlined in §1. For simplicity, we can take all errors equal (§1.2). Let us consider all parameters with their optimal values, except parameter k , with its value increased one time its uncertainty, $a_k + \epsilon(a_k)$. The corresponding model values are

$$y^{\text{mod}}(x_i; a_1, \dots, a_k + \epsilon(a_k), \dots, a_m) = y^{\text{mod}}(x_i; a_1, \dots, a_m) + f_k(x_i) \epsilon(a_k),$$

and the rms residual is

$$\begin{aligned} \epsilon_{\text{fit}}^2 &= \frac{1}{n} \sum_{i=1}^n [y_i^{\text{obs}} - y^{\text{mod}}(x_i) - f_k(x_i) \epsilon(a_k)]^2 \\ &= \frac{1}{n} \sum_{i=1}^n [y_i^{\text{obs}} - y^{\text{mod}}(x_i)]^2 - \frac{2\epsilon(a_k)}{n} \sum_{i=1}^n [y_i^{\text{obs}} - y^{\text{mod}}(x_i)] f_k(x_i) + \frac{\epsilon^2(a_k)}{n} \sum_{i=1}^n f_k^2(x_i). \end{aligned}$$

The second term is small because $\langle y_i^{\text{obs}} - y^{\text{mod}}(x_i) \rangle = 0$, and the observation errors are uncorrelated with the values of $f_k(x_i)$. Thus,

$$\epsilon_{\text{fit}}^2 \simeq \epsilon_{\text{fit},\text{min}}^2 + \frac{\epsilon^2(a_k)}{n} \sum_{i=1}^n f_k^2(x_i),$$

where $\epsilon_{\text{fit},\text{min}}$ is the minimum rms residual, for the optimal values of the parameters a_1, \dots, a_m . Now let us consider the value of $\epsilon(a_k) = C_{kk} = (J^t J)^{-1}_{kk}$. The diagonal terms of $J^t J$ are

$$J^t J = \begin{pmatrix} \sum_{i=1}^n f_1^2(x_i) & & \\ & \ddots & \\ & & \sum_{i=1}^n f_m^2(x_i) \end{pmatrix}.$$

If the model parameters are not strongly correlated, it is expected that the non-diagonal terms of the array are much smaller than the diagonal terms. In this case, the diagonal terms of the inverse array $C = (J^t J)^{-1}$ will be approximately

$$C \simeq \begin{pmatrix} 1/\sum_{i=1}^n f_1^2(x_i) & & \\ & \ddots & \\ & & 1/\sum_{i=1}^n f_m^2(x_i) \end{pmatrix}.$$

Thus, the uncertainty of parameter k is approximately

$$\epsilon^2(a_k) \simeq \frac{\epsilon^2}{\sum_{i=1}^n f_m^2(x_i)}.$$

Thus, the rms residual for $a_k + \epsilon(a_k)$ is

$$\epsilon_{\text{fit}}^2 \simeq \epsilon_{\text{fit},\text{min}}^2 + \frac{\epsilon^2}{n}$$

If the model fits reasonably the observations, the observation error ϵ will be similar to the rms residual $\epsilon_{\text{fit},\text{min}}$, and

$$\epsilon_{\text{fit}}^2 \simeq \left(1 + \frac{1}{n}\right) \epsilon_{\text{fit},\text{min}}^2.$$

In conclusion, the uncertainty in each parameter a_k can be estimated by finding the increment Δa_k that makes the rms residual to be $(1 + 1/n)^{1/2}$ higher than the value for the optimum set of parameters.

3.3 Numerical example 1. Linear regression

Fit of 30 points of a straight line $y = a_0 + a_1x$, in the interval $x \in (-5, 5)$, plus a Gaussian noise of mean 0 and standard deviation 1. The parameters used are $a_0 = 0.7$, $a_1 = 0.3$.

The values shown are the average of the values obtained for 1 and 100 trials by lsq fitting. Three values of the error of each parameter are shown,

Covariance Error: value obtained from the covariance array of the lsq fit.

Montecarlo Error: rms of the values obtained for the parameter (shown only for 100 trials).

$(1 + 1/n)^{1/2}$ Error: variation in each parameter necessary to increase the rms residual of the fit a factor of $(1 + 1/n)^{1/2}$.

	a_0	a_1
1 trial:		
Value	2.9537	0.7679
Covariance Error	0.1920	0.0664
$(1 + 1/n)^{1/2}$ Error	0.1920	0.0664
100 trials:		
Value	3.0019	0.6956
Covariance Error	0.1758	0.0608
Montecarlo Error	0.1941	0.0654
$(1 + 1/n)^{1/2}$ Error	0.1758	0.0608

As can be seen in this case, both the Montecarlo and the $(1 + 1/n)^{1/2}$ errors are good estimates of the parameter errors obtained from the covariance array. In particular, the $(1 + 1/n)^{1/2}$ error estimate is unbiased because $\langle x \rangle = 0$, and the parameters a_0 and a_1 are uncorrelated, so that the non-diagonal terms of the covariance array vanish. For a case with $\langle x \rangle \neq 0$, the $(1 + 1/n)^{1/2}$ error underestimates the parameter errors.

3.4 Numerical example 2. Gaussian fit

Fit of 40 points of a Gaussian $y = a_1 \exp(-4 \ln 2(x - a_3)^2/a_2^2)$, in the interval $x \in (0, 100)$, plus a Gaussian noise of mean 0 and standard deviation 1. The parameters used are $a_1 = 100$, $a_2 = 10$, $a_3 = 40$.

The explanation of the table is the same as that of last example.

	a_1	a_2	a_3
1 trial:			
Value	99.2704	10.1071	40.0478
Covariance Error	0.5565	0.0742	0.0315
Montecarlo Error	–	–	–
$(1 + 1/n)^{1/2}$ Error	0.5130	0.0608	0.0315
100 trials:			
Value	100.2204	9.9886	40.0045
Covariance Error	0.5968	0.0779	0.0331
Montecarlo Error	0.6932	0.0885	0.0304
$(1 + 1/n)^{1/2}$ Error	0.5518	0.0637	0.0331

In this example of iterative non-linear fit, both the Montecarlo and the $(1 + 1/n)^{1/2}$ error are good estimates of the parameter errors.