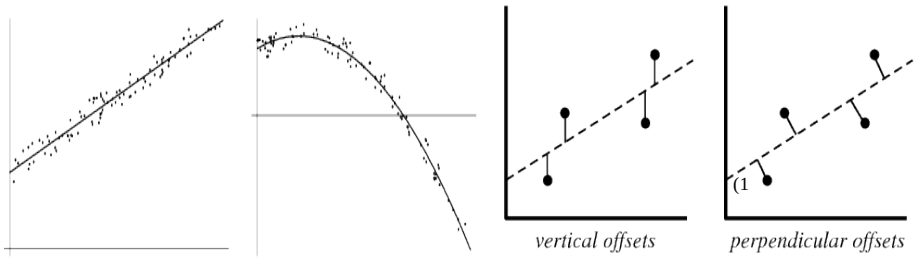
Least Squares Fitting

Least Square Fitting

- A mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets *("the residuals")* of the points from the curve.
- The sum of the *squares* of the offsets is used instead of the offset absolute values because this allows the residuals to be treated as a continuous differentiable quantity.
- However, because squares of the offsets are used, outlying points can have a disproportionate effect on the fit, a property which may or may not be desirable depending on the problem at hand.

Least Squares Method



Vertical least squares fitting proceeds by finding the sum of the *squares* of the *vertical* deviations R^2 of a set of *n* data points

$$R^2 \equiv \sum [y_i - f(x_i, a_1, a_2, \ldots, a_n)]^2$$

10.2 Non-linear least square fitting (NAM 6.10, H 6.6)

Example Fit the model function $F(t) = a + b\sin\omega(t - t_0)$ to experimental data.

t	0.5	0.8	1.0	1.2	1.5	1.8	2.0	2.4
y	0.3	0.3	0.5	0.9	1.4	1.1	0.5	0.3

This yields an **overdetermined non-linear system** (4 unknowns, 8 equations)

$$\begin{cases} a + b\sin\omega(0.5 - t_0) \approx 0.3 \\ a + b\sin\omega(0.8 - t_0) \approx 0.3 \\ a + b\sin\omega(1.0 - t_0) \approx 0.5 \\ a + b\sin\omega(1.2 - t_0) \approx 0.9 \\ a + b\sin\omega(1.5 - t_0) \approx 1.4 \\ a + b\sin\omega(1.8 - t_0) \approx 1.1 \\ a + b\sin\omega(2.0 - t_0) \approx 0.5 \\ a + b\sin\omega(2.4 - t_0) \approx 0.3 \end{cases} \quad \text{or} \quad \begin{cases} a + b\sin\omega(0.5 - t_0) - 0.3 \approx 0 \\ a + b\sin\omega(0.8 - t_0) - 0.3 \approx 0 \\ a + b\sin\omega(1.0 - t_0) - 0.5 \approx 0 \\ a + b\sin\omega(1.2 - t_0) - 0.9 \approx 0 \\ a + b\sin\omega(1.5 - t_0) - 1.4 \approx 0 \\ a + b\sin\omega(1.8 - t_0) - 1.1 \approx 0 \\ a + b\sin\omega(2.0 - t_0) - 0.5 \approx 0 \\ a + b\sin\omega(2.4 - t_0) - 0.3 \approx 0 \end{cases}$$

which can be solved by demanding that residuals between the measurements y_i and the model $F(t_i)$ be small $\|\mathbf{y} - \mathbf{F}\|_2 \approx 0$, i.e. $\|\mathbf{y} - \mathbf{F}\|_2 = \mathbf{f}(\mathbf{c}) \approx 0$.

Gauss-Newton solution obtained from $\mathbf{c}^{\text{new}} = \mathbf{c}^{\text{old}} + \delta\mathbf{c}, \quad \mathbf{J}^T\mathbf{J}\delta\mathbf{c} = -\mathbf{J}^T\mathbf{f}(\mathbf{c}^{\text{old}})$

The coefficients $\mathbf{c}$ are obtained from the Newton method for non-linear systems, with an increment $\delta\mathbf{c}$ solving the *overdetermined linear system* $\mathbf{J}\delta\mathbf{c} \approx -\mathbf{f}$. In Matlab, the solution of the normal equations can again be computed with `dc=-f\J`.

Least Squares Method

The square deviations from each point are therefore summed, and the resulting residual is then minimized to find the best fit line.

The condition for R^2 to be minimum is that $\frac{\partial (R^2)}{\partial a_i} = 0$ for $i = 1..n$

For a linear fit $f(a,b) = a + b\,x$, so $R^2(a,b) \equiv \sum_{i=1}^{n} [y_i - (a + b\,x_i)]^2$ and $\frac{\partial (R^2)}{\partial a} = -2 \sum_{i=1}^{n} [y_i - (a + b\,x_i)] = 0$

and $\frac{\partial (R^2)}{\partial b} = -2 \sum_{i=1}^{n} [y_i - (a + b\,x_i)]\, x_i = 0.$ These lead to the equations

$$n\,a + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i\, y_i .$$

In matrix forms:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix}, \quad \Rightarrow \quad \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix}.$$

Least Squares Method

The 2x2 matrix inverse is

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i \\ n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix},$$

$$a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\bar{y} \left(\sum_{i=1}^{n} x_i^2\right) - \bar{x} \sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}$$

$$b = \frac{n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\left(\sum_{i=1}^{n} x_i\, y_i\right) - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}$$

Least Squares Method

The previous formulas can be rewritten in a simpler form by defining the sums of squares

$$\mathrm{ss}_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2$$

$$\mathrm{ss}_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\sum_{i=1}^{n} y_i^2\right) - n\bar{y}^2$$

$$\mathrm{ss}_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y},$$

$$\sigma_x^2 = \frac{\mathrm{ss}_{xx}}{n}$$

$$\sigma_y^2 = \frac{\mathrm{ss}_{yy}}{n}$$

$$\mathrm{cov}(x, y) = \frac{\mathrm{ss}_{xy}}{n}.$$

σ_x^2 σ_y^2 : Variances

$\mathrm{cov}(x, y)$: Covariance

The overall quality of the fit is then parameterized in terms of a quantity known as the correlation coefficient, defined by

$$r^2 = \frac{\mathrm{ss}_{xy}^2}{\mathrm{ss}_{xx}\,\mathrm{ss}_{yy}},$$

The Meaning of *correlation coefficient*

- 0 indicates no linear relationship.
- +1 indicates a perfect positive linear relationship: as one variable increases in its values, the other variable also increases in its values via an exact linear rule.
- -1 indicates a perfect negative linear relationship: as one variable increases in its values, the other variable decreases in its values via an exact linear rule.
- Values between 0 and 0.3 (0 and -0.3) indicate a weak positive (negative) linear relationship via a shaky linear rule.
- Values between 0.3 and 0.7 (0.3 and -0.7) indicate a moderate positive (negative) linear relationship via a fuzzy-firm linear rule.
- Values between 0.7 and 1.0 (-0.7 and -1.0) indicate a strong positive (negative) linear relationship via a firm linear rule.
- The value of r squared is typically taken as “the percent of variation in one variable explained by the other variable,” or “the percent of variation shared between the two variables.”
- Linearity Assumption. The correlation coefficient requires that the underlying relationship between the two variables under consideration is linear. If the relationship is known to be linear, or the observed pattern between the two variables appears to be linear, then the correlation coefficient provides a reliable measure of the strength of the linear relationship. If the relationship is known to be nonlinear, or the observed pattern appears to be nonlinear, then the correlation coefficient is not useful, or at least questionable.

Non-Linear Least Square

Consider a set of m data points, $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, and a curve (model function) $y = f(x, \boldsymbol{\beta})$, that in addition to the variable x also depends on n parameters, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$, with $m \geq n$. It is desired to find the vector $\boldsymbol{\beta}$ of parameters such that the curve fits best the given data in the least squares sense, that is, the sum of squares

$$S = \sum_{i=1}^{m} r_i^2$$

is minimized, where the residuals (errors) r_i are given by

$$r_i = y_i - f(x_i, \boldsymbol{\beta})$$

for $i = 1, 2, \ldots, m$.

The minimum value of S occurs when the gradient is zero. Since the model contains *n* parameters there are *n* gradient equations:

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0 \quad (j = 1, \ldots, n).$$

In a non-linear system, the derivatives $\frac{\partial r_i}{\partial \beta_j}$ are functions of both the independent variable and the parameters, so these gradient equations do not have a closed solution. Instead, initial values must be chosen for the parameters. Then, the parameters are refined iteratively, that is, the values

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0 \quad (j = 1, \ldots, n).$$

In a non-linear system, the derivatives $\frac{\partial r_i}{\partial \beta_j}$ are functions of both the independent variable and the parameters, so these gradient equations are obtained by successive approximation,

$$\beta_j \approx \beta_j^{k+1} = \beta_j^k + \Delta\beta_j.$$

Here, *k* is an iteration number and the vector of increments, $\Delta\boldsymbol{\beta}$ is known as the shift vector.

At each iteration the model is linearized by approximation to a first-order Taylor series expansion about $\boldsymbol{\beta}^k$

$$f(x_i, \boldsymbol{\beta}) \approx f(x_i, \boldsymbol{\beta}^k) + \sum_j \frac{\partial f(x_i, \boldsymbol{\beta}^k)}{\partial \beta_j} \left(\beta_j - \beta_j^k\right) \approx f(x_i, \boldsymbol{\beta}^k) + \sum_j J_{ij} \, \Delta\beta_j.$$

Non-Linear Least Square

The Jacobian, **J**, is a function of constants, the independent variable *and* the parameters, so it changes from one iteration to the next.

Thus, in terms of the linearized model, $\frac{\partial r_i}{\partial \beta_j} = -J_{ij}$ and the residuals are given by

$$r_i = \Delta y_i - \sum_{s=1}^{n} J_{is} \, \Delta\beta_s; \; \Delta y_i = y_i - f(x_i, \boldsymbol{\beta}^k).$$

Substituting these expressions into the gradient equations, they become

$$-2 \sum_{i=1}^{m} J_{ij} \left(\Delta y_i - \sum_{s=1}^{n} J_{is} \, \Delta\beta_s \right) = 0$$

which, on rearrangement, become *n* simultaneous linear equations, the **normal equations**

$$\sum_{i=1}^{m} \sum_{s=1}^{n} J_{ij} J_{is} \, \Delta\beta_s = \sum_{i=1}^{m} J_{ij} \, \Delta y_i \qquad (j = 1, \ldots, n).$$

The normal equations are written in matrix notation as

$$\left(\mathbf{J^T J}\right) \boldsymbol{\Delta\beta} = \mathbf{J^T} \, \boldsymbol{\Delta y}.$$

When the observations are not equally reliable, a weighted sum of squares may be minimized,

$$S = \sum_{i=1}^{m} W_{ii} r_i^2.$$

Each element of the diagonal weight matrix **W** should, ideally, be equal to the reciprocal of the error variance of the measurement.

The normal equations are then

$$\left(\mathbf{J^T W J}\right) \boldsymbol{\Delta\beta} = \mathbf{J^T W} \, \boldsymbol{\Delta y}.$$

These equations form the basis for the Gauss–Newton algorithm for a non-linear least squares problem.

Geometric Interpretation & Convergence Criteria

In linear least squares the objective function, *S*, is a quadratic function of the parameters.

$$S = \sum_i W_{ii} \left(y_i - \sum_j X_{ij}\beta_j \right)^2$$

When there is only one parameter the graph of *S* with respect to that parameter will be a parabola.
the objective function is quadratic with respect to the parameters only in a region close to its minimum value, where

$$S \approx \sum_i W_{ii} \left(y_i - \sum_j J_{ij}\beta_j \right)^2$$

The more the parameter values differ from their optimal values, the more the contours deviate from elliptical shape.
A consequence of this is that initial parameter estimates should be as close as practicable to their (unknown!) optimal values.
It also explains how divergence can come about as the Gauss–Newton algorithm is convergent only when the objective function is approximately quadratic in the parameters.

The common sense criterion for convergence is that the sum of squares does not decrease from one iteration to the next. However this criterion is often difficult to implement in practice, for various reasons. A useful convergence criterion is

$$\left| \frac{S^k - S^{k+1}}{S^k} \right| < 0.0001.$$

The value 0.0001 is somewhat arbitrary and may need to be changed. In particular it may need to be increased when experimental errors are large.
An alternative criterion is

$$\left| \frac{\Delta\beta_j}{\beta_j} \right| < 0.001, \qquad j = 1, \ldots, n.$$

Again, the numerical value is somewhat arbitrary; 0.001 is equivalent to specifying that each parameter should be refined to 0.1% precision. This is reasonable when it is less than the largest relative standard deviation on the parameters.