Least Squares Estimation

The general objective of least squares estimation is that you want to minimize some measure of the deviation between the model estimate $\tilde{f}(x_i)$ and the known values $f_o(x_i)$ - thus hoping that you will have a "good fit" of the model to the data. The closeness of the fit will depend on data errors and the selected function model.

We will show at a later time that if these deviations $\{ \tilde{f}(x_i) - f_o(x_i) \}$ are Gaussian (normally distributed), the "best estimate" is found by minimizing the sum of squares of the deviations; i.e., we minimize

$$E = \sum_{i=1}^{N} [\tilde{f}(x_i) - f_o(x_i)]^2 \qquad (1)$$

where N is the number of observations.

Our first example will be a one-dimensional case, in which the model is a 2nd-order polynomial. Thus the model is

$$\tilde{f}(x) = a_0 + a_1 x + a_2 x^2 \qquad (2)$$

and therefore we minimize

$$E = \sum_{i=1}^{N} [(a_0 + a_1 x_i + a_2 x^2_i) - f_o(x_i)]^2 \qquad (3)$$

Note that for this quadratic model, N must be equal to or greater than 3.

Since the x_i and $f_o(x_i)$ are known, we need to determine a_0, a_1 and a_2.

To find a_0, a_1 and a_2 such that E is a minimum, we derive the **normal equations** by setting

$$\frac{\partial E}{\partial a_0} \rightarrow 0, \quad \frac{\partial E}{\partial a_1} \rightarrow 0, \quad \frac{\partial E}{\partial a_2} \rightarrow 0$$

Note that setting $\frac{\partial A}{\partial (\ )} \rightarrow 0$ usually finds an extrema only, but since E is positive definite, this extrema is always a minima in this case.

Thus, performing these operations on eq. (3) yields

$$\frac{\partial E}{\partial a_0} = 2 \Sigma [a_0 + a_1 x_i + a_2 x^2_i - f_o(x_i)] \qquad (4)$$

$$\frac{\partial E}{\partial a_1} = 2 \Sigma [a_0 + a_1 x_i + a_2 x^2_i - f_o(x_i)] x_i \qquad (5)$$

$$\frac{\partial E}{\partial a_2} = 2 \Sigma [a_0 + a_1 x_i + a_2 x^2_i - f_o(x_i)] x^2_i \qquad (6)$$

Taking the summation inside eq. (4) (and dividing out the 2), we have

$$\sum_{i=1}^{N} a_0 + \sum_i a_1 x_i + \sum_i a_2 x^2_i - \sum_i f_o(x_i) = 0$$

or

$$a_0 N + a_1 \sum_i x_i + a_2 \sum_i x^2_i - \sum_i f_o(x_i) = 0$$

and, dividing by N and defining the ensemble average $\overline{(\ )} = 1/N \sum_{i=1}^{N} (\ )$

we have $a_0 + a_1 \overline{x_i} + a_2 \overline{x^2_i} = \overline{f_o(x_i)}$

Similarly, eqs. (5) and (6) can be rewritten as

$$a_0 \overline{x_i} + a_1 \overline{x^2_i} + a_2 \overline{x^3_i} = \overline{x_i f_o(x_i)}$$

$$a_0 \overline{x^2_i} + a_1 \overline{x^3_i} + a_2 \overline{x^4_i} = \overline{x^2_i f_o(x_i)}$$

These normal equations can be written in matrix form:

$$\begin{bmatrix} 1 & \overline{x_i} & \overline{x_i^2} \\ \overline{x_i} & \overline{x_i^2} & \overline{x_i^3} \\ \overline{x_i^2} & \overline{x_i^3} & \overline{x_i^4} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \overline{f_o(x_i)} \\ \overline{x_i f_o(x_i)} \\ \overline{x_i^2 f_o(x_i)} \end{bmatrix} \qquad (7)$$

or $\underset{\sim}{X}\ \underset{\sim}{A} = \underset{\sim}{F}$ (8)

Note that the coefficient matrix $\underset{\sim}{X}$ is symmetric.

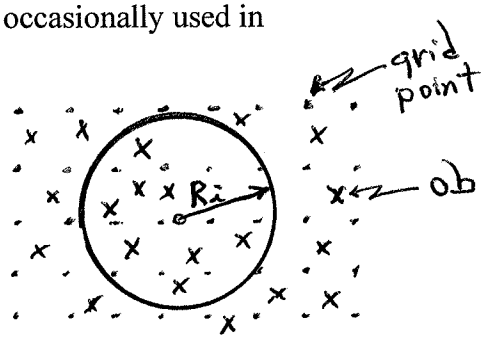
The solution of (8) is $\underset{\sim}{A} = \underset{\sim}{X}^{-1} \underset{\sim}{F}$ provided the inverse exists (i.e. $\underset{\sim}{X}$ is nonsingular.)

Note: Moments in $\underset{\sim}{X}$ increase greatly for higher-order polynomials; $\underset{\sim}{X}$ may be singular or ill-conditioned beyond 5th or 6th order.

Now we are ready to examine Sec. 2.1 of Daley, which is an example of <u>two-dimensional local polynomial fitting</u>. This technique, developed by Panofsky (1949), produced the first objective analysis ever used in NWP. Local fitting is still occasionally used in regional data analysis studies.

The problem is illustrated by the diagram on the right, for which we need to determine an analysis value $f_A(x,y)$ at the center point r_i of the circle with an <u>influence radius</u> R_i. Only observations within the influence radius are used to determine $f_A(x,y)$; - i.e.

$$x^2_k + y^2_k \leq R^2_i$$



Denote the observations $f_o(x_k,y_k)$, $k = 1(1) K$, where K is the no. of observations in the <u>influence region</u>.

Define $x = y = 0$ at analysis point r_i.

Assume $f_A(x,y)$ can be represented by the following 2-D polynomial expansion:

$$f_A(x,y) = \sum_m \sum_n c_{m,n}\, x^m\, y^n \qquad m, n \geq 0,\ m+n \leq M \qquad (1)$$

Note that at the center point $x = 0$, $y = 0$, in eq. (1) yields $f_A(0,0) = c_{00}$.

At each r_i, we want to minimize the differences between the function model (1) and the observations; i.e., we minimize I, given by

$$I = \tfrac{1}{2} \sum_{k=1}^{K_i} \left[\sum_m \sum_n c_{m,n}\, x^m{}_k\, y^n{}_k - f_o(x_k,y_k) \right]^2 \qquad (2)$$

As usual, we minimize I by taking the first derivative of (2) w.r.t. the unknown coefficients $c_{m,n}$ and setting the result = 0:

$$\frac{\partial I}{\partial c_{m,n}} = \sum_{k=1}^{K_i} x^m{}_k\, y^n{}_k \left[\sum_\mu \sum_\nu c_{\mu,\nu}\, x^\mu{}_k\, y^\nu{}_k - f_o(x_k,y_k) \right] \longrightarrow 0$$

Which can be rewritten as (class exercise)

$$\sum_\mu \sum_\nu c_{\mu,\nu} \sum_{k=1}^{K_i} x_k{}^{m+\mu}\, y_k{}^{n+\nu} = \sum_{k=1}^{K_i} x^m{}_k\, y^n{}_k\, f_o(x_k,y_k) \qquad (3)$$

Now, defining the ensemble average here as $\overline{(\ )} = 1/K_i \sum_{k=1}^{K_i} (\ )$, also show that (3) can be written in matrix form given by eq. (2.1.4) in Daley when

M = 2. (This permits the following coefficients: $c_{00}, c_{10}, c_{20}, c_{11}, c_{01}, c_{02}$, and the matrices will be of order

$$(6 \times 6)(6 \times 1) = (6 \times 1) \qquad \text{(2.1.4 - Daley)}$$

Equation (2.1.4) is then solved for the 6 coefficients to obtain one analysis value. We then move to the next grid point and repeat the process; i.e. – local fitting.

Least squares estimation derivation

Since least squares minimization is used or implied throughout this course, we will examine it more closely. Here we will show that the least squares minimization value, S_a, is a maximum likelihood estimate of S.

Consider N observations $s_1, s_2, s_3, \ldots\ldots S_N$ of a variable s - at one point.

Assume each observation is taken with a different instrument and the observational error of each ob. is $\varepsilon_n = s_n - s$, n = 1(1)N, where s is the (unknown) true value.

Assume errors are random, unbiased, and normally distributed, and recall the formula for the normal (Gaussian) probability density function p_G given by

$$p_G(s) = \frac{1}{\sigma_s \sqrt{2\pi}} \exp\left[-\frac{(s - \langle s \rangle)^2}{2\sigma_s^2}\right] \qquad (1)$$

where σ_s is the standard deviation of s and

σ_s^2 is the variance (2^{nd} moment) of s, denoted as $\langle (s - \langle s \rangle)^2 \rangle$

where $\langle s \rangle$ is the expected value or mean of s given by

$$\langle s \rangle = \int_{-\infty}^{\infty} s\, p(s)\, ds$$

(i.e. – the sum over all values of s, each one weighted by its probability.)

p(s) is any probability density function (pdf) that describes the distribution of s. A pdf gives the probability that s lies between s and s + ds, and satisfies the properties:

$$p(s) \geq 0; \qquad \int_{-\infty}^{\infty} p(s)\, ds = 1$$

[For a review of statistics, see Appendix D in Daley.]

So, if the errors are Gaussian, we use eq. (1) and the definition of ε_n to write that the probability that the error in the nth observation lies between ε_n and $\varepsilon_n + d\varepsilon_n$ is given by

$$p(\varepsilon_n) = \frac{1}{\sigma_n \sqrt{2\pi}} \exp\left[-\frac{\varepsilon_n^2}{2\sigma_n^2} \right] \qquad (2)$$

where

$$\sigma_n^2 = \langle (s_n - s)^2 \rangle = \langle \varepsilon_n^2 \rangle = \int_{-\infty}^{\infty} \varepsilon_n^2 \, p(\varepsilon_n) \, d\varepsilon_n \qquad (3)$$

Note that in (2) we have used the fact that $\langle \varepsilon_n \rangle = 0$, since errors are unbiased.

Now, for N = 1, we have one observation s_1 with error ε_1. Thus the most probable value of s is <u>that value for which $p(\varepsilon_1)$ is a maximum</u>.

Since $\langle \varepsilon_n \rangle = 0$, this occurs at $\varepsilon_1 = 0$, or, via the left side of (3), when $s_1 = s$. (Note that these results come from the properties of a Gaussian distribution.)

When there are N > 1 obs., we need to compute the <u>joint probability</u> that

ε_1 lies between $\varepsilon_1 + d\varepsilon_1$, ε_2 lies between $\varepsilon_2 + d\varepsilon_2$, ..., ε_N lies between $\varepsilon_N + d\varepsilon_N$ is the product of all the probabilities – which we write as

$p(\varepsilon_1, \varepsilon_2, \ldots\ldots, \varepsilon_N) = p(\varepsilon_1)\, p(\varepsilon_2) \ldots\ldots p(\varepsilon_N)$, which, via eq. (2) is

$$p(\varepsilon_1, \varepsilon_2 \ldots\ldots, \varepsilon_N) = \prod_{n=1}^{N} \frac{1}{\sigma_n \sqrt{2\pi}} \exp\left(-\frac{\varepsilon_n^2}{2\sigma_n^2} \right)$$

which can also be written

$$= \left[\prod_{n=1}^{N} \frac{1}{\sigma_n \sqrt{2\pi}} \right] \exp\left[-\sum_{n=1}^{N} \frac{(s_n - s)^2}{2\sigma_n^2} \right] \qquad (4)$$

since exponents are additive.

In this formulation, the most probable value of s (s_a) is that value for which the probability

$p(\varepsilon_1, \varepsilon_2, \ldots\ldots, \varepsilon_N)$ is a maximum. (That is, if we know the most likely ε_n, that that will direct us to the most likely s_n.)

This will occur when the quantity $\sum_{n=1}^{N} \frac{(s_n - s)}{2\sigma_n^2}$ in (4) is a <u>minimum</u> (so

that the exponent is large).

Thus the most probable value, s_a, often called the maximum likelihood estimate of s, must minimize

$$I = \tfrac{1}{2} \sum_{n=1}^{N} \sigma_n^{-2} (s_a - s_n)^2 \qquad (5)$$

As before, to minimize a quantity, we differentiate I w.r.t. the unknown quantity s_a and set to zero:

$$\sum_{n=1}^{N} \sigma_n^{-2} (s_a - s_n) = \frac{s_a - s_1}{\sigma_1^2} + \frac{s_a - s_2}{\sigma_2^2} + \cdots \frac{s_a - s_n}{\sigma_N^2} = 0$$

or

$$s_a \sum_{n=1}^{N} \sigma_n^{-2} - \sum_{n=1}^{N} \sigma_n^{-2} s_n = 0$$

and, solving for s_a, we have

$$s_a = \frac{\sum_{n=1}^{N} \sigma_n^{-2} s_n}{\sum_{n=1}^{N} \sigma_n^{-2}} \qquad (6)$$

Therefore, the most probable value, or most likelihood estimate is given by a weighted sum of the observations, with weights inversely proportional to the observational error variance.

Once we have an analysis estimate s_a, it is important to examine the error ε_a of the analysis - we will usually want to know the expected error variance of the analysis.

Thus we define $\varepsilon_a = s_a - s$ (7)

And, using eqs. (6) and (7), we can write the expected error variance of the estimate as

$$\langle \varepsilon_a^2 \rangle = \left\langle \left[\frac{\sum_{n=1}^{N} \sigma_n^{-2} (s_n - s)}{\sum_{n=1}^{N} \sigma_n^{-2}} \right]^2 \right\rangle \qquad (8)$$

Class exercise: Show that (8) can be simplified to

$$\langle \varepsilon_a^2 \rangle = \left[\sum_{n=1}^{N} \sigma_n^{-2} \right]^{-1} \qquad (9)$$

(Hint: Do N = 2 case first before do general case.)

Now suppose all the obs. were taken with the same instrument. Thus, $\sigma_n^2 = \sigma^2$ for all n, and eqs. (6) and (9) reduce to

$$s_a = 1/N \sum_{n=1}^{N} s_n \qquad \text{and} \qquad \langle \varepsilon_a^2 \rangle = \sigma^2 / N \qquad (10a,b)$$

Thus s_a (10a) becomes a simple ensemble average. It is important to note from (10b) that when $N > 1$, $\varepsilon_a^2 < \sigma^2$ which means that the analysis value is better than any one observation.