LEAST-SQUARES FIT (Chapter 8)

Fit the 'best' straight line (parabola, etc.) to a given set of points. This will be done by minimizing the sum of squares of the *vertical* distances (called residuals) from the points to the line.

Normal Equations

Let us assume that we want to fit a *parabola,* whose general equation is $y = a + b\,x + c\,x^2$.

Minimize (by varying a, b and c):

$$S \equiv \sum_{i=1}^{n} \left(y_i - a - b\,x_i - c\,x_i^2\right)^2$$

The three partial derivatives, i.e. $\frac{\partial S}{\partial a}$, $\frac{\partial S}{\partial b}$ and $\frac{\partial S}{\partial c}$, must be equal to zero:

$$-2\sum_{i=1}^{n} \left(y_i - a - b\,x_i - c\,x_i^2\right) \;=\; 0$$

$$-2\sum_{i=1}^{n} \left(y_i - a - b\,x_i - c\,x_i^2\right) x_i \;=\; 0$$

$$-2\sum_{i=1}^{n} \left(y_i - a - b\,x_i - c\,x_i^2\right) x_i^2 \;=\; 0$$

Cancelling 2 and reorganizing yields:

$$
\begin{aligned}
a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i \\
a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} x_i y_i \\
a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} x_i^2 y_i
\end{aligned}
$$

or (in matrix form):

$$
\begin{bmatrix}
n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\
\sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\
\sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4
\end{bmatrix}
\begin{bmatrix} a \\ b \\ c \end{bmatrix}
=
\begin{bmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} x_i y_i \\
\sum_{i=1}^{n} x_i^2 y_i
\end{bmatrix}
$$

For straight line:

$$
\begin{bmatrix}
n & \sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2
\end{bmatrix}
\begin{bmatrix} a \\ b \end{bmatrix}
=
\begin{bmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} x_i y_i
\end{bmatrix}
$$

For a cubic:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 & \sum_{i=1}^{n} x_i^6 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \\ \sum_{i=1}^{n} x_i^3 y_i \end{bmatrix}$$

etc.

The smallest value of S, say S_{min}, is (going back to the 'parabola' case):

$$\sum_{i=1}^{n} y_i^2 - \begin{bmatrix} a & b & c \end{bmatrix} \cdot \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \end{bmatrix}$$

This can serve to define the so called typical error:

$$\sqrt{\frac{S_{\text{min}}}{n-3}}$$

where 3 is the number of coefficients in the fitted parabola (2 for a straight line, 4 for a cubic, etc.).

Example:

Fit a parabola to the following data:

x:	1	2	3	4	5
y:	2	3	2	2	1

Normal equations:

5	15	55	10
15	55	225	27
55	225	979	89

We solve these by the so called

Gaussian Elimination

followed by backward substitution.

There are four elementary operations which can be used to modify a linear set of equations *without changing its solution*. These are:

1. Adding (subtracting) a multiple of a row to (from) any other row.
2. Interchanging two rows (i.e. equations).
3. Multiplying a row (equation) by a *non-zero* number.
4. Interchanging two columns of the coefficient matrix (we have to keep track of these).

Proceeding in a systematic way, with the help of the first *two* elementary operations we can make all elements below the main diagonal equal to zero. It is

done in this order:

×	×	×	×
1	×	×	×
2	4	×	×
3	5	6	×

How do we deal with zero *on* the main diagonal?

Example:

(Our normal equations). Subtract the first row multiplied by 3 (11) from the second (third) row:

5	15	55	10
0	10	60	–3
0	60	374	–21

Subtract the second row multiplied by 6 from the third row:

5	15	55	10
0	10	60	–3
0	0	14	–3

Starting from the bottom, we now find the unknowns by *backward substitution:*

$$c = -\frac{3}{14}$$

$$b = \frac{-3 + 60 \cdot \frac{3}{14}}{10} = \frac{69}{70}$$

$$a = \frac{10 - 15 \cdot \frac{69}{70} + 55 \cdot \frac{3}{14}}{5} = \frac{7}{5}$$

(verify!). *Typical error* of the fit:

$$\sqrt{\frac{22 - 10 \cdot \frac{7}{5} - 27 \cdot \frac{69}{70} + 89 \cdot \frac{3}{14}}{5-3}} = 0.4781$$

Symmetric Data

When x-values are spaced symmetrically around a center (say $\hat{x}$), the computation simplifies. Using the same example, we fit

$$y = \hat{a} + \hat{b}(x - \hat{x}) + \hat{c}(x - \hat{x})^2$$

where $\widehat{x} =$ 3. The normal equations now read

$$\begin{bmatrix} n & \sum_{i=1}^{n}(x_i - \hat{x}) & \sum_{i=1}^{n}(x_i - \hat{x})^2 \\ \sum_{i=1}^{n}(x_i - \hat{x}) & \sum_{i=1}^{n}(x_i - \hat{x})^2 & \sum_{i=1}^{n}(x_i - \hat{x})^3 \\ \sum_{i=1}^{n}(x_i - \hat{x})^2 & \sum_{i=1}^{n}(x_i - \hat{x})^3 & \sum_{i=1}^{n}(x_i - \hat{x})^4 \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(x_i - \hat{x})\, y_i \\ \sum_{i=1}^{n}(x_i - \hat{x})^2 y_i \end{bmatrix}$$

But. due to symmetry, the sum of any *odd* power of $x_i - \hat{x}$ yields 0:

$$\begin{bmatrix} n & 0 & \sum_{i=1}^{n}(x_i-\hat{x})^2 \\ 0 & \sum_{i=1}^{n}(x_i-\hat{x})^2 & 0 \\ \sum_{i=1}^{n}(x_i-\hat{x})^2 & 0 & \sum_{i=1}^{n}(x_i-\hat{x})^4 \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(x_i-\hat{x})\, y_i \\ \sum_{i=1}^{n}(x_i-\hat{x})^2 y_i \end{bmatrix}$$

which means we can separate the '*odd*' coefficients

$$\begin{bmatrix} n & \sum_{i=1}^{n}(x_i-\hat{x})^2 \\ \sum_{i=1}^{n}(x_i-\hat{x})^2 & \sum_{i=1}^{n}(x_i-\hat{x})^4 \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(x_i-\hat{x})^2 y_i \end{bmatrix}$$

from '*even*'

$$\hat{b}\sum_{i=1}^{n}(x_i-\hat{x})^2 = \sum_{i=1}^{n}(x_i-\hat{x})\, y_i$$

Example:

Using the old data and $\hat{x} = 3$ yields

$x - \hat{x}$:	–2	–1	0	1	2
y:	2	3	2	2	1

Equations for $\hat{a}$ and $\hat{c}$:

$$\begin{bmatrix} 5 & 10 \\ 10 & 34 \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} 10 \\ 17 \end{bmatrix}$$

solved by

$$\begin{bmatrix} 34 & -10 \\ -10 & 5 \end{bmatrix} \begin{bmatrix} 10 \\ 17 \end{bmatrix} \div 70 = \begin{bmatrix} \frac{17}{7} \\ -\frac{3}{14} \end{bmatrix}$$

The $\hat{b}$ equation is trivial: $10\hat{b} = -3 \Rightarrow \hat{b} = -\frac{3}{10}$.

The best fitting parabola is:

$$y = \frac{17}{7} - \frac{3}{10}(x-3) - \frac{3}{14}(x-3)^2$$

(the same as before, when expanded). It also yields the same typical error:

$$\sqrt{\frac{22 - 10 \cdot \frac{17}{7} + 17 \cdot \frac{3}{14} - 3 \cdot \frac{3}{10}}{5-3}} = 0.4781$$

Weighted Fit

Some of the x-y pairs may be repeated. We can then simplify the table by adding a new row of weights:

x:	1	2	3	4	5
y:	2	3	2	2	1
w:	4	2	1	1	3

Now, we minimize

$$S \equiv \sum_{i=1}^{n} w_i \left(y_i - a - b\,x_i - c\,x_i^2\right)^2$$

(each squared residual multiplied by weight).

Normal equations:

$$\begin{bmatrix} \sum_{i=1}^{n} w_i & \sum_{i=1}^{n} w_i x_i & \sum_{i=1}^{n} w_i x_i^2 \\ \sum_{i=1}^{n} w_i x_i & \sum_{i=1}^{n} w_i x_i^2 & \sum_{i=1}^{n} w_i x_i^3 \\ \sum_{i=1}^{n} w_i x_i^2 & \sum_{i=1}^{n} w_i x_i^3 & \sum_{i=1}^{n} w_i x_i^4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} w_i y_i \\ \sum_{i=1}^{n} w_i x_i y_i \\ \sum_{i=1}^{n} w_i x_i^2 y_i \end{bmatrix}$$

solved as before.

The symmetry trick will work now only if the w_i 'cooperate'.

S_{min} is now computed by

$$\sum_{i=1}^{n} w_i y_i^2 - \begin{bmatrix} a & b & c \end{bmatrix} \cdot \begin{bmatrix} \sum_{i=1}^{n} w_i y_i \\ \sum_{i=1}^{n} w_i x_i y_i \\ \sum_{i=1}^{n} w_i x_i^2 y_i \end{bmatrix}$$

Typical error:

$$\sqrt{\frac{n}{\sum_{i=1}^{n} w_i} \cdot \frac{S_{\text{min}}}{n-3}}$$

Example:

We fit the best *straight line* to the previous data.

Normal equations:

11	30	21
30	112	49

Solution:

$$\begin{bmatrix} 112 & -30 \\ -30 & 11 \end{bmatrix} \begin{bmatrix} 21 \\ 49 \end{bmatrix} \div 332 = \begin{bmatrix} \frac{441}{166} \\ -\frac{91}{332} \end{bmatrix}$$

i.e.

$$y = \frac{441}{166} - \frac{91}{332} x$$

Typical error:

$$\sqrt{\frac{5}{11}\cdot\frac{45-21\cdot\frac{441}{166}+49\cdot\frac{91}{332}}{5-2}}=0.6326$$

Linear Models

We can replace x, x^2,.... of our polynomial model by any other *specific* functions of x (e.g. $\frac{1}{x}$, e^x, $\ln x$, etc.),

$$y = a\,f_1(x) + b\,f_2(x) + c\,f_3(x)$$

in general.

Normal equations:

$$\begin{bmatrix} \sum_{i=1}^{n} f_1(x_i)^2 & \sum_{i=1}^{n} f_1(x_i)f_2(x_i) & \sum_{i=1}^{n} f_1(x_i)f_3(x_i) \\ \sum_{i=1}^{n} f_2(x_i)f_1(x_i) & \sum_{i=1}^{n} f_2(x_i)^2 & \sum_{i=1}^{n} f_2(x_i)f_3(x_i) \\ \sum_{i=1}^{n} f_3(x_i)f_1(x_i) & \sum_{i=1}^{n} f_3(x_i)f_2(x_i) & \sum_{i=1}^{n} f_3(x_i)^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} f_1(x_i)\,y_i \\ \sum_{i=1}^{n} f_2(x_i)\,y_i \\ \sum_{i=1}^{n} f_3(x_i)\,y_i \end{bmatrix}$$

This time, symmetry is of no help.

Example:

Let us fit our old data (without weights) by

$$y = \frac{a}{x} + b + c\,x^3$$

i.e. $f_1(x) = \frac{1}{x}$, $f_2(x) = 1$ and $f_3(x) = x^3$.

Normal equations:

$$\left.\begin{array}{ccc|c} \sum_{i=1}^{5} \frac{1}{x_i^2} & \sum_{i=1}^{5} \frac{1}{x_i} & \sum_{i=1}^{5} x_i^2 & \sum_{i=1}^{5} \frac{y_i}{x_i} \\ \sum_{i=1}^{5} \frac{1}{x_i} & \sum_{i=1}^{5} 1 & \sum_{i=1}^{5} x_i^3 & \sum_{i=1}^{5} y_i \\ \sum_{i=1}^{5} x_i^2 & \sum_{i=1}^{5} x_i^3 & \sum_{i=1}^{5} x_i^6 & \sum_{i=1}^{5} y_i x_i^3 \end{array}\right.$$

or, numerically

$$\left.\begin{array}{ccc|c} \frac{5269}{3600} & \frac{137}{60} & 55 & \frac{73}{15} \\ \frac{137}{60} & 5 & 225 & 10 \\ 55 & 225 & 20515 & 333 \end{array}\right.$$

Gaussian (forward) elimination results in

$$\left.\begin{array}{ccc|c} \frac{5269}{3600} & \frac{137}{60} & 55 & \frac{73}{15} \\ 0 & \frac{7576}{5269} & \frac{66675}{479} & \frac{12686}{5269} \\ 0 & 0 & \frac{37673515}{7576} & -\frac{314271}{3788} \end{array}\right.$$

back substitution yields: $c = -\frac{628542}{37673515}$, $b = \frac{2253323}{684973}$ and $a = -\frac{8891100}{7534703}$

Typical error:

$$\sqrt{\frac{22 + \frac{8891100}{7534703} \cdot \frac{73}{15} - \frac{2253323}{684973} \cdot 10 + \frac{628542}{37673515} \cdot 333}{5-3}} = 0.4483$$

We could easily modify this procedure to incorporate weights.

Our last example is a clear indication that *exact* computation (using fractions) is getting rather tedious (or even impossible), and that we should consider switching to decimals. This necessitates modifying the Gaussian-elimination procedure, to keep the round off errors of the computation in check..