"Best" Line through points

Given a pair of points in the plane there is a unique line that joins them. How about if we are given three points? Clearly, if the points are not collinear, then we cannot ask for a line that joins these three points. The problem appears to become even worse if we ask for a line that fits more points, but we "know" that more data should be more information!

So, given a bunch of points (x_i, y_i) in the plane, we are asking for a line that fits these points. One way to approach this problem is to say that the points (x_i, y_i) contain experimental errors and *actually* lie on a line and our job is to determine this line.

A line in the plane is given by an equation of the form $y - mx = c$. For each fixed m, we calculate $y_i - mx_i = c_i$. If m is the slope of the line we are looking for, then these values c_i are of the form $c + e_i$, where e_i is a measure of the experimental error in the measurement of the pair (x_i, y_i). As usual, we assume that this experimental error e_i is normally distributed around 0 with standard deviation s (which is independent of i).

In that case, the likelihood (density) of obtaining the result that we have is (here N is the number of points)

$$L = \prod_i^N \frac{\exp(-e_i^2/2s)}{s\sqrt{2\pi}}$$

Equivalently, the log-likelihood is given by

$$l = -(N/2)\log(2\pi) - N\log(s) - \sum_i^N \frac{e_i^2}{2s}$$

Since s can be assumed to be a quantity that is determined by the experimental setup, it is "fixed". Thus l is maximum if $\sum_i^N e_i^2$ is *minimum*.

In other words, the *maximum likelihood estimate* for the parameters m and c (which determine the line) is associated with the case where the sum of squares of the errors is *least*. We note that this is under the assumption that the experimental errors distributed normally with mean 0 and some standard deviation s that is fixed.

This is one way to derive the method of *least square estimation* which defines the estimator as the one that minimises the sum of the squares of the errors.

Returning to the original problem, we can formulate it as follows. Consider the vectors $\mathbf{y} = (y_1, \ldots, y_N)$, $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{u} = (1, \ldots, 1)$. We are looking for constants m and c so that $\mathbf{e} = \mathbf{y} - m\mathbf{x} - c\mathbf{u}$ is of the least length. This is solved by "dropping a perpendicular" from the vector $\mathbf{y}$ to the plane spanned by $\mathbf{x}$ and $\mathbf{u}$. The base of the perpendicular is $m\mathbf{x} + c\mathbf{u}$ and the length of the perpendicular is the length of $\mathbf{e}$.

Since $\mathbf{e}$ is perpendicular to $\mathbf{x}$ and $\mathbf{u}$ and we have $\mathbf{y} = m\mathbf{x} + c\mathbf{u} + \mathbf{e}$, we obtain the linear equations:

$$\mathbf{x} \cdot \mathbf{y} = m\mathbf{x} \cdot \mathbf{x} + c\mathbf{x} \cdot \mathbf{u}$$
$$\mathbf{u} \cdot \mathbf{y} = m\mathbf{x} \cdot \mathbf{u} + c\mathbf{u} \cdot \mathbf{u}$$

We can solve these equations to obtain m and c (providing $\mathbf{x}$ and $\mathbf{u}$ are linearly independent; which is the case if some $x_i \neq x_j$).

"Best" Linear Fit

The above situation can easily be generalised as follows.

We make a sequence of measurements that produce tuples of the form $(x_{i,1}, \ldots, x_{i,r}, y_i)$. Theory leads us to believe that these satisfy an equation of the form $y = m_1 x_1 + \cdots + m_r x_r + c$. As usual, we know that there will be experimental errors so our actual equations look like

$$y_i = m_1 x_{i,1} + \cdots + m_r x_{i,r} + c + e_i$$

where e_i denotes an experimental error that follows a normal distribution $N(0, s)$ for some s which is a consequence of the experimental setup (in particular, is independent of i). Moreover, we can either assume that the measurements for different i are independent, so that e_i are independent random variables or, at the very least that these are uncorrelated random variables. In that case, as in the 2-dimensional case, we can compute the log-likelihood as

$$l = -(N/2)\log(2\pi) - N\log(s) - \sum_{i}^{N} \frac{e_i^2}{2s}$$

where N is the number of tuples as above. Again, we see that this is maximised when the length of the vector $\mathbf{e} = (e_1, \ldots, e_N)$ is minimised. In other words, the least squares estimator is the same as the maximum likelihood estimator.

Solving this problem can again be posed as a problem in geometry by considering the vectors

$$\mathbf{y} = (y_1, y_2, \ldots, y_N)$$
$$\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{N,j})$$
$$\mathbf{u} = (1, 1, \ldots, 1)$$

Our equation then becomes

$$\mathbf{y} = m_1 \mathbf{x}_1 + \cdots + m_r \mathbf{x}_r + c\mathbf{u} + \mathbf{e}$$

In the optimal case $\mathbf{e}$ will be *perpendicular* to each of the vectors $\mathbf{x}_i$ and the vector $\mathbf{u}$. Thus, we can obtain the optimal values $m_1, \ldots, m_N$ and c by solving the system of linear equations

$$\begin{aligned}
\mathbf{x}_1 \cdot \mathbf{y} &= m_1 \mathbf{x}_1 \cdot \mathbf{x}_1 + \cdots + m_r \mathbf{x}_1 \cdot \mathbf{x}_r + c \mathbf{x}_1 \cdot \mathbf{u} \\
&\vdots \\
\mathbf{x}_r \cdot \mathbf{y} &= m_1 \mathbf{x}_r \cdot \mathbf{x}_1 + \cdots + m_r \mathbf{x}_r \cdot \mathbf{x}_r + c \mathbf{x}_r \cdot \mathbf{u} \\
\mathbf{u} \cdot \mathbf{y} &= m_1 \mathbf{u} \cdot \mathbf{x}_1 + \cdots + m_r \mathbf{u} \cdot \mathbf{x}_r + c \mathbf{u} \cdot \mathbf{u}
\end{aligned}$$

This is a system of $r+1$ linear equations in $r+1$ unknowns, which can be solved under the assumption that the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_r$ and $\mathbf{u}$ are linearly independent. (If not, then we can eliminate one of the sets of "independent" variables $\mathbf{x}_j$ from consideration.)

This solution gives the least square fit for the dependent variable y or, equivalently least square estimator for the quantities $m_1, \ldots, m_r$ and c.

Whither non-linear functions?

The above may leave the impression that we are not considering the situation where y is a non-linear function of x_i's. However, that is not the case!

Suppose we expect y to be a function $f(z_1, \ldots, z_q; m_1, \ldots, m_r)$ where f is non-linear in the variables z_i, but is *linear* in the *parameters* m_k. For example, f is a polynomial function in the z_i's and m_k are the undetermined coefficients of the polynomial. Or f is a linear combination of sine and cosine functions in the z_i's and m_k are the (Fourier-type) coefficients that we are trying to determine. In each of these cases, we can re-write the function f in the form of a linear combination $\sum_i m_i f_i(\mathbf{z})$. We can then put $x_i = f_i(\mathbf{z})$ and reduce the problem to that described above.

The combinations of sines and cosines is the context is where Gauss discovered the method given above while trying to determine the orbit of Ceres; he simultaneously discovered the Fast Fourier Transform which is a quick way to carry out the calculation.

The assumptions

The description of the problem and its solution by means of linear equations is dependent on certain assumptions:

Exogeneity The assumption that all the experimental errors are captured in e_i. In other words, even though the $x_{i,j}$ are measured quantities, there are no errors mixed up in them.

Independence The errors e_i are independent random variables that are normally distributed around 0.

Common variance This is sometimes also called *homoskedasticity*. This is the condition that the random variables e_i all have the same variance s.

Lack of perfect multicollinearity Since the term "independence" could be confusing in this context, we don't use it! However, this is the condition that the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_r, \mathbf{u}$ are *linearly* independent.

Linearity This is the condition that y depends linearly on $m_1, \ldots, m_r$.

BLUE

The solution to the problem above gives an estimator for the tuple $(m_1, \ldots, m_r, c)$. This is sometimes called Best Linear Unbiased Estimator (given the acronym BLUE).

We say that the estimator is *Linear* because it has the form $\mathbf{C} \cdot \mathbf{y}$ for a suitable matrix $\mathbf{C}$ whose entries are (non-linear) functions of the $\mathbf{x}_{i,j}$. The estimator is only linear in $\mathbf{y}$.

We say that the estimator is *Unbiased* because its *expected* value is the tuple $(m_1, \ldots, m_r, c)$ which gives the precise linear expression for y in terms of $x_1, \ldots, x_r$. One calculates that this means that $\mathbf{C} \cdot \mathbf{x}_i = (0, \ldots, 1, \ldots, 0)$ (where 1 occurs in the i-th place) and $\mathbf{C} \cdot \mathbf{u} = (0, \ldots, 1)$.

Given any estimate $(n_1, \ldots, n_r, d)$ for required tuple, we define the associated *residual* as the difference

$$\tilde{\mathbf{r}} = \mathbf{y} - (n_1\mathbf{x}_1 + \cdots + n_r\mathbf{x}_r + d\mathbf{u})$$

The length of the residual represents how far our estimate fails to match the experimental result. Thus, one possible notion for the *Best* estimate would be one for which the residual has the *smallest* length among all possible estimates. We can also interpret this in terms of log-likelihood being the maximum as seen earlier.

While calculating $\mathbf{C}$ is possibly, it should be pointed out that this can be computationally intensive and there *are* quicker methods to compute the estimated tuple $(m_1, \ldots, m_r, c)$ directly.