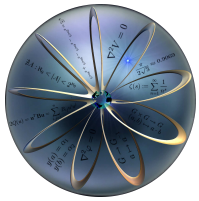

planetmath.org

Math for the people, by the people.

linear least squares fit

Canonical name	LinearLeastSquaresFit
Date of creation	2013-03-22 17:24:16
Last modified on	2013-03-22 17:24:16
Owner	rspuzio (6075)
Last modified by	rspuzio (6075)
Numerical id	13
Author	rspuzio (6075)
Entry type	Definition
Classification	msc 15-00
Related topic	RegressionModel
Related topic	GaussMarkovTheorem

One of the most common uses of least squares fitting is fitting a straight line to data. Whilst, in general, it is difficult to determine the curve which best fits the data, in this case there is a relatively simple formula which can be used.

Theorem 1. *Suppose we have a data set* $(x_1, y_1), \ldots, (x_n, y_n)$. *Then the straight line which best fits this set is given as*

$$y = \frac{ns - pq}{nr - p^2}x + \frac{qr - ps}{nr - p^2}$$

where

$$p = \sum_{k=1}^{n} x_k \tag{1}$$

$$q = \sum_{k=1}^{n} y_k \tag{2}$$

$$r = \sum_{k=1}^{n} x_k^2 \tag{3}$$

$$s = \sum_{k=1}^{n} x_k y_k \tag{4}$$

Proof. Being the best fitting line means minimizing the merit function M, given as

$$M(a, b) = \sum_{k=0}^{n} (ax_k + b - y_k)^2$$

with respect to the parameters a and b. Expanding the square, this can be written as

$$M(a, b) = ra^2 + 2pab + nb^2 - 2sa - 2qb + t$$

where p, q, r, s are as above and

$$t = \sum_{k=1}^{n} y_k^2.$$

This function M is a quadratic polynomial; moreover, from its definition as a sum of squares, it is clear that the highest order terms are positive

definite, hence it has a minimum and all that remains is to find that minimum. To do this, we set the derivatives equal to zero to obtain the following equations:

$$0 = \frac{\partial M(a,b)}{\partial a} = 2ar + 2pb - 2s \tag{5}$$

$$0 = \frac{\partial M(a,b)}{\partial b} = 2pa + 2nb - 2q \tag{6}$$

These equations are easily solved to give

$$a = \frac{ns - pq}{nr - p^2} \tag{7}$$

$$b = \frac{qr - ps}{nr - p^2}; \tag{8}$$

substituting in the equation $y = ax + b$ for a straight line, we obtain the answer given above. □

Because of the ease with which one can make a least squares fit of a line, this technique is often adapted to fitting other sorts of curves by making a change of variables. Two common cases of this practice are power laws and exponentials.

Suppose that one wants to fit some data to a curve of the form $y = ce^{kx}$. Making a change of variable $y = e^u$ and defining $b = \log c$, the equation of the curve becomes $u = kx + b$. One can therefore fit the data set $(x_1, \log y_1), \ldots (x_n, \log y_n)$ to a straight line.

Suppose that one wants to fit some data to a curve of the form $y = cx^p$. Making a change of variable $x = e^v$, $y = e^u$ and defining $b = \log c$, the equation of the curve becomes $u = pv + b$. One can therefore fit the data set $(\log x_1, \log y_1), \ldots (\log x_n, \log y_n)$ to a straight line.

Although convenient and common, this procedure can be a cheat because changing variables and making a least squares fit of a line is not the same as making a least squares fit to a curve. The reason for this is that the merit functions are different and will not, in general have a minimum in the same place. However, if the data happen to approximately lie on a power curve or an exponential, then the answer obtained by changing variables and fitting will be an approximation to the correct answer. Depending on what one is doing, this approximation may be good enough or one may use it as a starting point for some algorithm to compute the correct minimum.