Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Orthogonal linear regression is used when there are two variables, X and Y, which both have error of measurement, and/or natural variation due to sampling from a population. Because of this there is no sense in which one variable can be regarded as an independent variable, and the other a dependent variable with added noise: they are really covariates, but they could be sufficiently related to justify fitting a straight line. The problem is to decide the criteria to use when selecting a best-fit line.

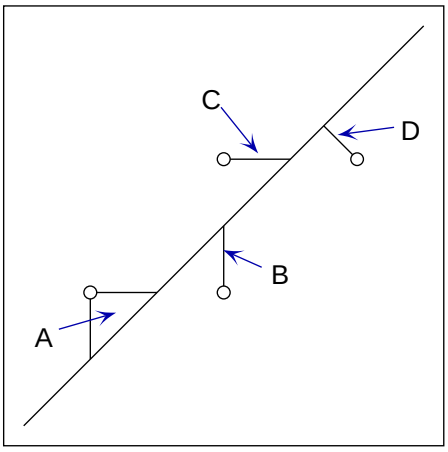
From the SIMFIT main menu choose [A/Z], open program **linfit**, choose one of the options for orthogonal or reduced major axis regression and inspect the default test file `swarm.tf1` which has the following data.

x	y
-5.0754	6.4669
-0.1053	11.6754
3.4949	15.4471
3.9864	5.0136
5.1110	7.8573
5.4251	-0.1269
5.7351	2.5006
5.9965	12.8566
6.5293	13.0522
6.6922	11.7522
6.7427	7.9817
8.9142	6.0645
9.6825	19.2638
12.0221	14.5156
14.5866	19.9856
15.4610	17.7134
16.3355	22.4164
16.8102	9.7428
16.9810	23.3692
17.2585	17.3129
18.9608	5.2797
20.2275	16.5656
24.2327	26.7548
25.0702	12.7738
27.6169	25.1028

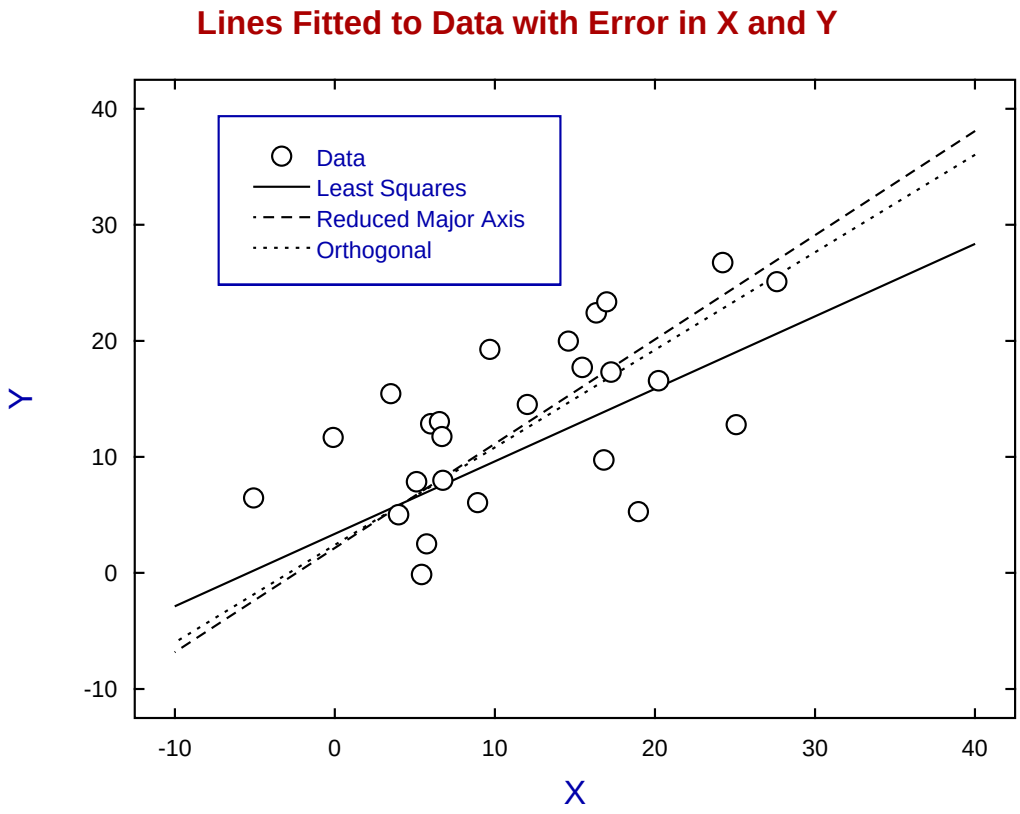
The following straight line procedures could be considered.

1. Least squares fit for $y = ax + b$, i.e. $Y(X)$.
2. Least squares fit for $x = \alpha y + \beta$, i.e. $X(Y)$.
3. Reduced major axis regression.
   This minimizes the sum of the areas of the triangles formed by projecting across and up or down from the data points to the best-fit line.
4. Orthogonal or major axis regression.
   This minimizes the sum of squares of the orthogonal projections from the points to the best-fit line.

Reduced major axis regression minimizes the sum of the areas of the triangles **A**, least squares $Y(X)$ minimizes the sum of squares of the lengths **B**, least squares $X(Y)$ minimizes the sum of squares of lengths **C**, while orthogonal regression, minimizes the sum of squares of the lengths **D** as in the next diagram.



The following plot shows the fit of lines by least squares, reduced major axis, and major axis (orthogonal) regression to the data in test file `swarm.tf1`. In general it seems that if the X and Y data are similarly scaled then the choice depends on the variance of X and Y. If the variance of Y is very much greater than the variance of X then least squares regression of Y on X could be preferred, and when the variance of Y is very much less than that of X then least squares regression of X on Y might be better. With similar variance in Y and Y, as in correlation analysis, then either both linear regression lines should be plotted, or one of the alternatives described in this tutorial should be used if only one line is to be plotted.



Theory

For n pairs (x_i, y_i) with mean $x = \bar{x}$ and mean $y = \bar{y}$, the variances and covariance required are

$$S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$S_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Also, for an arbitrary point (x_i, y_i) and a straight line defined by $y = a + bx$ the squares of the vertical, horizontal, and orthogonal (i.e. perpendicular) distances, v_i^2, h_i^2, and o_i^2 between the point and the line are

$$v_i^2 = [y_i - (a + bx_i)]^2$$
$$h_i^2 = v_i^2/b^2$$
$$o_i^2 = v_i^2/(1 + b^2).$$

Ordinary least squares

If x is regarded as an exact variable free from random variation or measurement error while y has random variation, then the best fit line from minimizing the sum of v_i^2 is

$$y_1(x) = \hat{\beta}_1 x + [\bar{y} - \hat{\beta}_1 \bar{x}]$$

where $\hat{\beta}_1 = S_{xy}/S_{xx}$. However, if y is regarded as an exact variable while x has random variation, then the best fit line for x as a function of y from minimizing the sum of h_i^2 would be

$$x_2(y) = (1/\hat{\beta}_2)y + [\bar{x} - (1/\hat{\beta}_2)\bar{y}]$$

where $\hat{\beta}_2 = S_{yy}/S_{xy}$ or, rearranging to express the line as $y_2(x)$,

$$y_2(x) = \hat{\beta}_2 x + [\bar{y} - \hat{\beta}_2 \bar{x}],$$

emphasizing that the slope of the regression line for $y_2(x)$ is the reciprocal of the slope for $x_2(y)$. Since neither of these two best fit lines can be regarded as satisfactory, SimFIT plots both lines such that $y_1(x)$ covers the range of x values while $x_2(y)$ covers the range of y values. However these two lines intersect at $(\bar{x}, \bar{y})$ and, from the fact that the ratio of slopes equals the square of the correlation coefficient, that is,

$$r^2 = \hat{\beta}_1/\hat{\beta}_2,$$

then two best fit lines with similar slopes suggests strong linear correlation, whereas one line almost parallel to the x axis and the other almost parallel to the y axis would indicate negligible linear correlation. For instance, if there is no linear correlation between x and y, then the slope of the regression line for $y(x)$ i.e. $\hat{\beta}_1$ would be zero, as would be the slope of the regression line for $x(y)$ i.e. $1/\hat{\beta}_2$ leading to $r^2 = 0$. Conversely strong linear correlation would lead to $\hat{\beta}_1 = \hat{\beta}_2$ and $r^2 = 1$.

The major axis and reduced major axis lines to be discussed next are attempts to get round the necessity to plot two lines and just have one best fit line intermediate between these two lines to represent the correlation.

The major axis line

Here it is the sum of o_i^2, the squares of the orthogonal distances between the points and the best fit line, that is minimized to yield the slope as

$$\hat{\beta}_3 = \frac{1}{2}\left(\hat{\beta}_2 - (1/\hat{\beta}_1) + \gamma\sqrt{4 + (\hat{\beta}_2 - (1/\hat{\beta}_1))^2}\right)$$

where $\gamma = 1$ if $S_{xy} > 0$, $\gamma = 0$ if $S_{xy} = 0$, and $\gamma = -1$ if $S_{xy} < 0$, so that the major axis line is

$$y_3(x) = \hat{\beta}_3 x + [\bar{y} - \hat{\beta}_3 \bar{x}].$$

Actually $\hat{\beta}_3$ is the slope of the first principal component axis and so it points in the direction of maximum variability.

The reduced major axis line

Instead of minimizing the sum of squares of the vertical distances v_i^2, or horizontal distances h_i^2, it is possible to minimize the sum of the areas of the triangles formed by the v_i, h_i with the best fit line as hypotenuse, i.e. $v_i h_i/2$, to obtain the reduced major axis line as

$$y_4(x) = \hat{\beta}_4 x + [\bar{y} - \hat{\beta}_4 \bar{x}].$$

Here

$$\begin{aligned}\hat{\beta}_4 &= \gamma\sqrt{S_{yy}/S_{xx}} \\ &= \gamma\sqrt{\hat{\beta}_1\hat{\beta}_2}\end{aligned}$$

so that the slope of the reduced major axis line is the geometric mean of the slopes of the regression of y on x and x on y.

Weighting

In the unlikely case that weighting of one of the set is observations is desired, then the variable to be weighted would have to be specified as the Y variable, and weighted fitting could then performed using program **qnfit**.