

LSQ fit of a straight line

Robert Estalella

2009 December

1 Least-squares fit of a straight line with two free parameters

1.1 Standard linear regression, error only in y

This is the standard linear regression fit. Let us assume that we have n points in the x - y plane, (x_i, y_i) , $i = 1, \dots, n$. The values x_i are supposed to be exact, while the error associated with each point y_i is ϵ_i . We are searching the straight line

$$y = px + q,$$

depending on two free parameters, the slope p and the intercept q (the value of y for $x = 0$), which best fits the points. Since it can be shown that the regression line passes through the average point $(\langle x \rangle, \langle y \rangle)$, its expression can be given as

$$y - \langle y \rangle = p(x - \langle x \rangle),$$

with a value of the intercept q

$$q = \langle y \rangle - p\langle x \rangle.$$

The best fit minimizes the objective function

$$Q = \sum_{i=1}^n \frac{(y_i - px_i - q)^2}{\epsilon_i^2}.$$

The value found for p is given by the standard formula,

$$p = \frac{m_{xy}}{\sigma_x^2},$$

where, as usual,

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2, \quad \sigma_y^2 = \langle y^2 \rangle - \langle y \rangle^2,$$

and m_{xy} is the centered cross-moment,

$$m_{xy} = \langle xy \rangle - \langle x \rangle \langle y \rangle.$$

The averages are taken using $1/\epsilon_i^2$ as weights, that is,

$$\langle x^r y^s \rangle = \frac{\sum_{i=1}^n x_i^r y_i^s / \epsilon_i^2}{\sum_{i=1}^n 1 / \epsilon_i^2}.$$

The fit residual, i.e. the rms vertical distance from the straight line to the points is given by

$$\epsilon_{\text{fit}} = \sigma_y \sqrt{1 - r^2} = \sqrt{\sigma_y^2 - p^2 \sigma_x^2},$$

where r is the correlation coefficient,

$$r = \frac{m_{xy}}{\sigma_x \sigma_y}.$$

The uncertainty in the slope of the regression line is given by

$$\epsilon_p = \frac{1}{\sqrt{n}} \frac{\sigma_y}{\sigma_x} \sqrt{1 - r^2} = \frac{1}{\sqrt{n}} \sqrt{\frac{\sigma_y^2}{\sigma_x^2} - p^2},$$

and that of the intercept,

$$\epsilon_q = \frac{\sigma_y}{\sqrt{n}} \sqrt{1 - r^2} \sqrt{1 + \frac{\langle x \rangle^2}{\sigma_x^2}} = \frac{1}{\sqrt{n}} \sqrt{\sigma_y^2 - p^2 \sigma_x^2} \sqrt{1 + \frac{\langle x \rangle^2}{\sigma_x^2}}.$$

1.2 Linear regression, error in x and y

Let us assume that there are errors in both x and y . The errors in x and y are assumed to be equal, $\epsilon(x_i) = \epsilon(y_i) = \epsilon$. If not, both axes have to be re-scaled in order to have equal errors. The best-fit straight line is the line that minimizes the weighted sum of distances d_i of the points to the line,

$$Q = \sum_{i=1}^n \frac{d_i^2}{\epsilon^2}.$$

For each point (x_i, y_i) , the distance d_i to the straight line is given by

$$d_i^2 = \frac{[(y_i - \langle y \rangle) - p(x_i - \langle x \rangle)]^2}{1 + p^2},$$

while the distance from $(\langle x \rangle, \langle y \rangle)$ to the projection of the point (x_i, y_i) onto the line is given by

$$l_i^2 = \frac{[(x_i - \langle x \rangle) + p(y_i - \langle y \rangle)]^2}{1 + p^2}.$$

The value of the slope p is

$$p = c \pm \sqrt{1 + c^2} = c + \text{sign}(m_{xy}) \sqrt{1 + c^2},$$

where c is given by

$$c = \frac{\sigma_y^2 - \sigma_x^2}{2m_{xy}}.$$

The fit residual, ϵ_{fit} , i.e. the rms residual distance to the regression line is given by

$$\epsilon_{\text{fit}}^2 = \epsilon_d^2 = \frac{1}{n} \sum_{i=1}^n d_i^2 = \frac{\sigma_y^2 + p^2 \sigma_x^2 - 2p m_{xy}}{1 + p^2}.$$

A similar expression is obtained for the rms length along the regression line,

$$\epsilon_l^2 = \frac{1}{n} \sum_{i=1}^n l_i^2 = \frac{\sigma_x^2 + p^2 \sigma_y^2 + 2p m_{xy}}{1 + p^2}.$$

The equivalent length (exact for a uniform distribution along the regression line), can be taken as $2\sqrt{3}\epsilon_l$.

The uncertainty in the slope of the regression line, in terms of the fit residual, is given by

$$\epsilon_p = \frac{p}{\sqrt{n}} \frac{\sqrt{\sigma_x^2 + \sigma_y^2}}{m_{xy}} \epsilon_{\text{fit}},$$

and that of the intercept, in terms of the fit residual and the slope uncertainty, by

$$\epsilon_q^2 = \frac{1 + p^2}{n} \epsilon_{\text{fit}}^2 + \langle x \rangle^2 \epsilon_p^2.$$

2 Least-squares fit of a straight line passing through a fixed point

2.1 Line passing through the origin, error only in y

The problem is to find the straight line passing through a fixed point. Without loss of generality, the fixed point can be taken as the origin $(0, 0)$,

$$y = p x.$$

If it is not the origin, a simple translation of coordinates can make it to be the origin. (The problem is formally similar to the general linear regression, when the averages $\langle x \rangle$ and $\langle y \rangle$ are zero.)

The line searched is the line that best fits n points (x_i, y_i) , $i = 1, \dots, n$. The values x_i are supposed to be exact, while the error associated with each value y_i is ϵ_i . The value of the single free parameter, the slope p , is the value that minimizes the objective function

$$Q = \sum_{i=1}^n \frac{(y_i - p x_i)^2}{\epsilon_i^2}.$$

The value of p can be easily calculated and is given by

$$p = \frac{\langle xy \rangle}{\langle x^2 \rangle}.$$

The fit residual, i.e. the rms vertical distance of the points to the line is given by

$$\epsilon_{\text{fit}} = \sqrt{\langle y^2 \rangle - p^2 \langle x^2 \rangle}.$$

The uncertainty in the slope p is given by

$$\epsilon_p = \frac{1}{\sqrt{n}} \sqrt{\frac{\langle y^2 \rangle}{\langle x^2 \rangle} - p^2}.$$

2.2 Line passing through the origin, error in x and y

The problem is the same as the latter, but now let us assume that there are errors in both x and y . The errors in x and y are assumed to be equal, $\epsilon(x_i) = \epsilon(y_i) = \epsilon$. If not, both axes have to be re-scaled in order to have equal errors.

The value of the single free parameter, the slope p , is the value that minimizes the objective function

$$Q = \sum_{i=1}^n \frac{d_i^2}{\epsilon^2},$$

where d_i is the distance of point (x_i, y_i) to the straight line $y = px$, given by

$$d_i^2 = \frac{(y_i - px_i)^2}{1 + p^2}.$$

The value of p can be calculated and is given by

$$p = c \pm \sqrt{1 + c^2} = c + \text{sign}(\langle xy \rangle) \sqrt{1 + c^2},$$

and c is given by

$$c = \frac{\langle y^2 \rangle - \langle x^2 \rangle}{2\langle xy \rangle}.$$

The fit residual (the rms distance to the line) is given by

$$\epsilon_{\text{fit}}^2 = \epsilon_d^2 = \frac{1}{n} \sum_{i=1}^n d_i^2 = \frac{\langle y^2 \rangle + p^2 \langle x^2 \rangle - 2p \langle xy \rangle}{1 + p^2}.$$

The uncertainty in the slope, in terms of the fit residual, is given by

$$\epsilon_p = \frac{p}{\sqrt{n}} \frac{\sqrt{\langle x^2 \rangle + \langle y^2 \rangle}}{\langle xy \rangle} \epsilon_{\text{fit}},$$