Auxiliary Guide

The standard deviation of the inclination of a straight line

Consider the affine function relating a predictive (or independent) variable x to the response (or dependent) variable y

$$y = ax + b$$

with a and b real constants. When using the Least Squares Method (LSM) to fit a and b to a dataset $\{(x_i, y_i), i = 1..N\}$ where the standard deviation of the response variable y_i is σ_i , the variance of a (the variance is the square of the standard deviation) can be calculated as [4]

$$\sigma_a{}^2 = \frac{\sum_1^{\mathrm{n}} \frac{1}{\sigma_{\mathrm{i}}^2}}{\sum_1^n \frac{1}{\sigma_{\mathrm{i}}^2} \sum_1^n \frac{x_i^2}{\sigma_{\mathrm{i}}^2} - \left(\sum_1^n \frac{x_i}{\sigma_{\mathrm{i}}^2}\right)^2}$$

When all $\sigma_i = \sigma$, this expression reduces to

$$\sigma_a{}^2 = \frac{N\,\sigma^2}{N \sum_1^n x_i^2 - (\sum_1^n x_i)^2} \qquad (E1)$$

In order to obtain a simpler formula, we define a central coordinate x_c

$$x_c = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \text{(E2)}$$

and relative coordinates δ_i

$$\delta_i = x_i - x_c \qquad \text{(E3)}$$

The denominator of eq. (E1), after substituting both definitions, reduces to

$$N \sum_{i=1}^{N} (\delta_i + x_c)^2 - N^2 x_c{}^2 = N \sum_{i=1}^{N} \delta_i{}^2 + 2N x_c \sum_{i=1}^{N} \delta_i = N \sum_{i=1}^{N} \delta_i{}^2$$

The first identity comes from the cancelation of the last term of the left member with one of the terms in the expansion of the sum of squares of $\delta_i + x_c$, and the second, from the fact that $\sum_{i=1}^{N} \delta_i = 0$. Using the obtained result in the denominator of eq. (E1), it reduces to

$$\sigma_a{}^2 = \frac{\sigma^2}{\sum_{i=1}^{N} \delta_i{}^2} \qquad \text{(E4)}$$

This result shows that the uncertainty on a depends only on the *uncertainty* of the data y (not on the coordinate values) and the dispersion of the *x*-values chosen for measurement.

An even more interesting expression can be obtained whenever the variable x is sampled uniformly, using a consistent interval Δx between adjacent measurements. For algebraic simplicity, we choose N odd, so we can express this set of data by

$$x_i = x_c + i\Delta x \quad \text{(E5)}$$

where i is an integer in the range $-\frac{N-1}{2} \leq i \leq \frac{N-1}{2}$. To make the algebraic manipulations easier, we define an integer $\nu = \frac{N-1}{2}$ which enters only the intermediate calculations. Using relation (E5) to evaluate the denominator of eq. (4), it follows

$$\sum_{i=1}^{N} {\delta_i}^2 = \Delta x^2 \sum_{i=-\nu}^{\nu} i^2 = \Delta x^2 \frac{1}{3} \nu(\nu+1)(2\nu+1) = \Delta x^2 \frac{1}{12}(N^3 - N)$$

For sufficiently big N, the last N in the parentheses can be ignored. Replacing the denominator of eq. (E4) by the resulting expression, it is obtained

$$ {\sigma_a}^2 \cong \frac{12\,\sigma^2}{N\,(N\,\Delta x)^2} \quad (E6)$$

Although expression (E6) solves the problem, it is interesting to highlight the role of the choice of measurement interval in the uncertainty in a. If x_o is the smallest observed value of x, the greatest value is

$$x_f = (N-1)\Delta x + x_o \quad \Leftrightarrow \quad x_f - x_o = (N-1)\Delta x \approx N\,\Delta x$$

a good approximation when N is large, which is often the case. Replacing this result in in formula (E6), gives

$$ {\sigma_a}^2 \cong \frac{12\,\sigma^2}{N\left(x_f - x_o\right)^2}$$

or

$$\sigma_a \approx \frac{\sigma}{x_f - x_0}\sqrt{\frac{12}{N}} \qquad \text{(E7)}$$

This expression shows clearly that the uncertainty in the inclination a depends only on the range of values x, the uncertainty in the response variable y , and in the number of observed points, not on the quality of the fit.

Reference: Draper, N. R., Smith, H., 1998. Applied Regression Analysis, 3rd Edition. Wiley, Hoboken, New Jersey.