Royal Holloway
University of London

Department of Physics

Estimating error bars from the variation of measured values about a fitted line

Often in data analysis one is presented with a set of measurements of a quantity y corresponding to different values of a control variable x, as shown in Fig. 1. Suppose the x values are known with negligible error but that the y values have some point-to-point variation. These variations reflect the random uncertainty in the y values and this is what we want to estimate from the data using the method of least squares. More information on the statistical formalism can be found in, for example, [1] and [2].

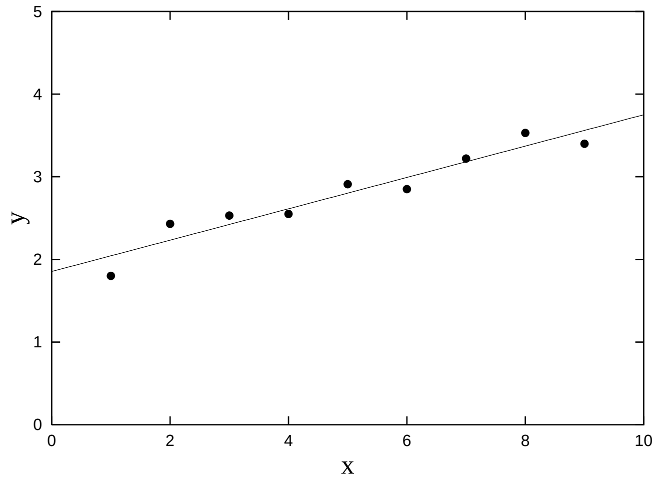

Figure 1: Measured values y at different values of the control variable x.

Often we can model the overall trend of the data as a polynomial such as a straight line, which we will use here as an example. That is, in the absence of measurement errors we would find $y = f(x; \mathbf{a})$ where

$$f(x; \mathbf{a}) = a_0 + a_1 x \; . \tag{1}$$

Here $\mathbf{a} = (a_0, a_1)$ is our vector of parameters describing the line.

Suppose we have data points (x_i, y_i) with $i = 1, \ldots, n$. To estimate the parameters $\mathbf{a}$ using the method of least squares, we should minimize the quantity

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{n} (y_i - f(x_i; \mathbf{a}))^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \; . \tag{2}$$

We therefore have to set the derivatives of χ^2 with respect to a_0 and a_1 equal to zero, i.e.,

$$\frac{\partial \chi^2}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = 0 , \tag{3}$$

$$\frac{\partial \chi^2}{\partial a_1} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) x_i = 0 . \tag{4}$$

Equations (3) and (4) can be written as

$$n\, a_0 + \sum_{i=1}^{n} x_i\, a_1 = \sum_{i=1}^{n} y_i , \tag{5}$$

$$\sum_{i=1}^{n} x_i\, a_0 + \sum_{i=1}^{n} x_i^2\, a_1 = \sum_{i=1}^{n} y_i x_i . \tag{6}$$

These are two linear equations for the two unknowns a_0 and a_1. They are of the form

$$A a_0 + B a_1 = C , \tag{7}$$

$$D a_0 + E a_1 = F , \tag{8}$$

where the definitions of A, B, etc., follow directly from comparison of (5) and (6) with (7) and (8). (Notice that in our example $B = D$.) The solutions are easily found to be

$$\hat{a}_0 = \frac{CE - BF}{AE - BD} , \tag{9}$$

$$\hat{a}_1 = \frac{AF - DC}{AE - BD} . \tag{10}$$

The solutions have been written with hats to emphasize that these are *estimators* for the true and in general unknown values a_0 and a_1.

If we had known the appropriate error bars (standard deviations) σ_i from the beginning, then we would have constructed the weighted chi-squared as

$$\chi^2_{\mathrm{w}} = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_i^2} . \tag{11}$$

Furthermore, if the hypothesis of a linear dependence is correct, then the expected value of the minimized χ^2_{w} is equal to the *number of degrees of freedom* of the fit, n_{dof}, which is the number of data points minus the number of fitted parameters. In our case we have $n_{\mathrm{dof}} = n - 2$. If we assume that the standard deviations are equal for all y values, i.e., $\sigma_i = \sigma$ for all i, then we expect

$$\sum_{i=1}^{n} \frac{(y_i - \hat{a}_0 - \hat{a}_1 x_i)^2}{\sigma^2} = \frac{\chi^2(\hat{\mathbf{a}})}{\sigma^2} = n - 2 \,, \tag{12}$$

where $\chi^2(\hat{\mathbf{a}})$ is the unweighted χ^2 from equation (2) evaluated with the estimates $\hat{a}_0$ and $\hat{a}_1$. So we can now estimate the standard deviation σ using

$$\hat{\sigma} = \sqrt{\frac{\chi^2(\hat{\mathbf{a}})}{n_{\rm dof}}} = \left[\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{a}_0 - \hat{a}_1 x_i)^2 \right]^{1/2} \,, \tag{13}$$

where $\hat{a}_0$ and $\hat{a}_1$ are obtained from equations (9) and (10). This represents the standard deviation needed to give a chi-square per degree of freedom of unity under the assumption of a linear relation for $f(x)$ and a common σ for all y_i.

Alternatively one might want to assume that all measured points have the same *relative* level of variation about their true values. That is, we could take the standard deviation σ_i to be

$$\sigma_i = f(x_i; \mathbf{a}) \varepsilon \approx y_i \varepsilon \,. \tag{14}$$

where ε represents the 'relative error' for each point. Requiring that chi-squared be equal to the number of degrees of freedom then means

$$\sum_{i=1}^{n} \frac{(y_i - \hat{a}_0 - \hat{a}_1 x_i)^2}{y_i^2 \varepsilon^2} = n - 2 \,. \tag{15}$$

The estimator for ε is therefore

$$\hat{\varepsilon} = \left[\frac{1}{n-2} \sum_{i=1}^{n} \frac{(y_i - \hat{a}_0 - \hat{a}_1 x_i)^2}{y_i^2} \right]^{1/2} \,. \tag{16}$$

It may not always be clear whether the assumption of constant σ or constant ε is valid. This must be determined from the data and from considerations of the origin of the random variation in the measured values.

References

[1] S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1997.

[2] G. Cowan, *Statistical Data Analysis*, Clarendon Press, Oxford, 1998.

G. Cowan, January, 2003