JagFit Error Method Guide

JagFit is a program designed to perform a least-squares fit to a set of data. One can perform a linear fit, exponential fit, power law fit, or a polynomial fit of up to 5^{th} order. For each fit type, the fit parameters are returned with errors.

When fitting a set of data, the user has three different ways of dealing with the errors in the measurements making up the data set.

1. The default method is the **Estimate dY** method. This method assumes that the error in a value of the independent variable x is much less than the error in the corresponding value of the dependent variable y and is therefore ignored. The error in y is unknown, but is assumed to be the same for all values of y in the set of data. With these assumptions, the fit parameter calculations do not depend on the error in the data. Once the fit parameters have been determined, the error in y can then be estimated by calculating the variance in y from the best-fit. Finally, the estimated uncertainties in y are used to in error propagation calculations to estimate uncertainties in the fir parameters.

2. The second method is the **Y Error Bars** method. This method also assumes that there is no error in the independent variable x, but it requires the user supply the error in the dependent variable y.

3. The third method is the **X and Y Error Bars** method. This method uses errors in both x and y as input by the user.

Table of Contents

Linear Fit

Given a set of data $\left(x_i, y_i\right)$ for which we believe the dependent variable y should depend linearly on independent variable x, we want the most probable values of a and b that give the best-fit line $y = a + bx$ for the set of data.

Y Error Bars Error Method
(See Chapter 6 of Bevington)

To determine the values a and b, we use the **method of maximum likelihood**. It is assumed that for each value of x_i, our measured value of y_i is drawn from a Gaussian distribution with a mean given by $y = a + bx$ and a standard deviation σ_{y_i}. Any error in x is assumed to be much smaller than the error in y and is therefore ignored. The most probable value for a and b then are the ones the minimize the 'goodness-of-fit' parameter χ^2 where $\chi^2 \equiv \sum_i \frac{(y_i - a - bx_i)^2}{\sigma_{y_i}^2}$. To minimize χ^2, we set $\frac{\partial \chi^2}{\partial a} = 0$ and $\frac{\partial \chi^2}{\partial b} = 0$. This gives us two equations in the two unknowns a and b. After a bit of algebra, the results are:

$$a = \frac{1}{D}\left(\sum \frac{x_i^2}{\sigma_{y_i}^2} \sum \frac{y_i}{\sigma_{y_i}^2} - \sum \frac{x_i}{\sigma_{y_i}^2} \sum \frac{x_i y_i}{\sigma_{y_i}^2} \right), \text{ and}$$

$$b = \frac{1}{D}\left(\sum \frac{1}{\sigma_{y_i}^2} \sum \frac{x_i y_i}{\sigma_{y_i}^2} - \sum \frac{x_i}{\sigma_{y_i}^2} \sum \frac{y_i}{\sigma_{y_i}^2} \right) \text{ where } D = \sum \frac{1}{\sigma_{y_i}^2} \sum \frac{x_i^2}{\sigma_{y_i}^2} - \left(\sum \frac{x_i}{\sigma_{y_i}^2} \right)^2$$

The error in the fit parameters (σ_a and σ_b) are found by propagating errors using the formula $\sigma_z^2 = \sum \sigma_{y_i}^2 \left(\frac{\partial z}{\partial y_i} \right)^2$ to obtain:

$$\sigma_a^2 = \sum \sigma_{y_i} \left(\frac{\partial a}{\partial y_i} \right)^2 \text{ which gives us } \sigma_a^2 = \frac{1}{D} \sum \frac{x_i^2}{\sigma_{y_i}^2} \text{ and}$$

$$\sigma_b^2 = \sum \sigma_{y_i} \left(\frac{\partial b}{\partial y_i} \right)^2 \text{ gives us } \sigma_b^2 = \frac{1}{D} \sum \frac{1}{\sigma_{y_i}^2}.$$

Estimate dY Error Method

(See Chapter 6 of Bevington)

In the *Y Error Bars Error Method* section above we outlined how the fit parameters and their errors are determined from a set of data for which we know the error in y and the error in x is ignored. If, however, the error in y is unknown, we can still fit the data if we assume that the error in y is the same for all values of y in the set of data.

With the simplifying assumption that $\sigma_{y_i} = \sigma_y$ for all y_i, then the above equations reduce to:

$$a = \frac{1}{E}\left(\sum x_i^2 \sum y_i^2 - \sum x_i \sum x_i y_i\right)$$

$$b = \frac{1}{E}\left(N\sum x_i y_i - \sum x_i \sum y_i\right)$$

with $\sigma_a^2 = \frac{\sigma_y^2}{E}\sum x_i^2$ and $\sigma_b^2 = \frac{N}{E}\sigma_y^2$ where $E = N\sum x_i^2 - \left(\sum x_i\right)^2$.

Note that with this simplification, neither a nor b depend on common error σ_y. Once a and b have been determined, an estimate for the common error σ_y can be made by calculating $\sigma_y^2 = \frac{1}{N-2}\left\{\sum\left(y_i - a - bx_i\right)^2\right\}$. This can be used to determine σ_a and σ_b.

X and Y Error Bars Error Method

In the *X and Y Error Bars* method we determine fit parameters of the best-fit line by using the error in x and y. We shall not provide the derivation of the following results in this help file. The interested reader and refer to Williamson's paper "Least-Squares fitting of a straight line,"

In this case, the 'goodness-of-fit' parameter is given by:

$$\chi^2 = \sum \frac{(x_i - x)^2}{\sigma_{x_i}^2} + \frac{(y_i - a - bx_i)^2}{\sigma_{y_i}^2}$$

This time, we must consider $\frac{\partial \chi^2}{\partial x_i} = 0$, $\frac{\partial \chi^2}{\partial a} = 0$ and $\frac{\partial \chi^2}{\partial b} = 0$.

This set of 3 equations can be solved for *b* to obtain

$$b = \frac{\sum w_i z_i y_i^{'}}{\sum w_i z_i x_i^{'}} \text{ where } w_i = \frac{1}{\sigma_{y_i}^2 + b^2 \sigma_{x_i}^2},$$

$x_i^{'} = x_i - \overline{x}$ and $\overline{x} = \frac{\sum w_i x_i}{\sum w_i}$ and $z_i = w_i(\sigma_{y_i}^2 x_i^{'} - b\sigma_{x_i}^2 y_i^{'})$ with similar definitions for $y_i^{'}$ and $\overline{y}$.

Note that the equation for *b* is solved iteratively. Once *b* has been determined we can calculate σ_b using:

$$\sigma_b = \frac{\sqrt{\sum w_i^2 \left\{ x_i^{'2} \sigma_{y_i}^2 + y_i^{'2} \sigma_{x_i}^2 \right\}}}{w_i \left\{ \frac{x_i^{'} y_i^{'}}{b} + 4 z_i^{'}(z_i - x_i^{'}) \right\}} \text{ where } z_i^{'} = z_i - \overline{z} \text{ and } \overline{z} = \frac{\sum w_i z_i}{\sum z_i}$$

Finally, $a = \overline{y} - b\overline{x}$ and σ_a is given by:

$$\sigma_a = \sqrt{\left(\sum w_i \right)^{-1} + \frac{2\left(\overline{x} + 2\overline{z}\right)\overline{z}}{\sum w_i \left\{ \frac{x_i^{'} y_i^{'}}{b} + 4 z_i^{'}(z_i - x_i^{'}) \right\}} + \left(\overline{x} + 2\overline{z}\right)^2 \sigma_b^2}\ .$$

Exponential Fit

Given a set data (x_i, y_i) for which we believe the dependent variable *y* should depend exponentially on independent variable *x*, we want the most probable values of *a* and *b* that give the best-fit to the function $y = ae^{bx}$. After a change of variable, the exponential fit becomes a linear fit and can be done as described in the *Linear Fit* section above.

For any of the three possible fit methods, we begin by taking the natural log of both sides of $y = ae^{bx}$, to get $\ln(y) = \ln(a) + bx$, or $y^{'} = a^{'} + bx$ where $y^{'} = \ln(y)$, and $a^{'} = \ln(a)$. If either the *Y Error Bars* or the *X & Y Error Bars*, error method is chosen, then the errors in *y* are also transformed via $\sigma_{y^{'}} = \frac{\sigma_y}{y}$. We then do the same least-squares fit to a straight line. This will give us the fit parameters $a^{'}$ and b as well as errors in the fit parameters $\sigma_{a^{'}}$ and σ_b. Finally, we return the values of $a, \sigma_a, b,$ and σ_b. Namely, $a = e^{a^{'}}$, $\sigma_a = a \cdot \sigma_{a^{'}}$.

Power Law Fit

Given a set data (x_i, y_i), we want the most probable values of *a* and *b* that give the best-fit to the function $y = ax^b$. After a change of variable, the power law fit becomes a linear fit and can be done as described in the *Linear Fit* section above.

For any of the three possible fit methods, we first take the log of both sides of $y = ax^b$, to obtain $\log(y) = \log(a) + b\log(x)$, or $y^{'} = a^{'} + bx^{'}$ where $y^{'} = \log(y)$, $a^{'} = \log(a)$, and $x^{'} = \log(x)$. If the error method is the *Y Error Bar* method, then the errors in *y* are also transformed via $\sigma_{y^{'}} = \frac{\log(e)}{y} \cdot \sigma_y$. If the error method is the *X and Y Error Bar* method, then the errors in *x* are also similarly transformed via $\sigma_{x^{'}} = \frac{\log(e)}{x} \cdot \sigma_x$. We then do the same least-squares fit to a straight line. This will give us the fit parameters $a^{'}$ and b as well as errors in the fit parameters $\sigma_{a^{'}}$ and σ_b. Finally, we return the values of $a, \sigma_a, b,$ and σ_b. Namely, $a = 10^{a^{'}}$, $\sigma_a = \frac{a}{\log(e)} \cdot \sigma_{a^{'}}$.

Polynomial Fit

Y Error Bars Error Method
(See Chapter 7 of Bevington)

Now suppose we want to fit our set of data with the polynomial
$y = a_o + a_1 x + a_2 x^2 + a_3 x^3 + \ldots$

The fit parameters $a_0, a_1, a_2, a_3, \ldots$ are determined by minimizing χ^2 where:
$\chi^2 = \sum \frac{(y_i - a_o - a_1 x_i - a_2 x_i^2 - a_3 x_i^3 - \ldots)^2}{\sigma_{y_i}^2}$. As before, we set $\frac{\partial \chi^2}{\partial a_o} = 0$, $\frac{\partial \chi^2}{\partial a_1} = 0$, $\frac{\partial \chi^2}{\partial a_2} = 0$, etc. This results in m+1 equations in m+1 unknowns for an m$^{\text{th}}$ order polynomial. We put this set of equations in matrix form and want to solve $\mathbf{A}\boldsymbol{\alpha} = \mathbf{B}$ for $\boldsymbol{\alpha}$ where:

$$\mathbf{A} = \begin{pmatrix} \sum \frac{1}{\sigma_{y_i}^2} & \sum \frac{x_i}{\sigma_{y_i}^2} & \cdots & \sum \frac{x_i^m}{\sigma_{y_i}^2} \\ \sum \frac{x_i}{\sigma_{y_i}^2} & \sum \frac{x_i^2}{\sigma_{y_i}^2} & \cdots & \sum \frac{x_i^{m+1}}{\sigma_{y_i}^2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum \frac{x_i^m}{\sigma_{y_i}^2} & \sum \frac{x_i^{m+1}}{\sigma_{y_i}^2} & \cdots & \sum \frac{x_i^{m+m}}{\sigma_{y_i}^2} \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} \sum \frac{y_i}{\sigma_{y_i}^2} \\ \sum \frac{x_i y_i}{\sigma_{y_i}^2} \\ \vdots \\ \sum \frac{x_i^m y_i}{\sigma_{y_i}^2} \end{pmatrix}, \text{ and } \boldsymbol{\alpha} = \begin{pmatrix} \mathbf{a_o} \\ \mathbf{a}_1 \\ \vdots \\ \mathbf{a_m} \end{pmatrix}.$$

Next, QR decomposition is used to solve for $\boldsymbol{\alpha}$. I did not write the code for the QR decomposition. I used the "TNT numerical package and JAMA QR algorithm" downloaded from the internet.

It turns out that the inverse of **A** is the "error matrix". The diagonal elements of $\mathbf{A}^{-1}$ are the variances in the fit parameters and the off diagonal elements are the covariances. I used Bevington's matrix inversion routine (Appendix E.3) to obtain $\mathbf{A}^{-1}$. The square roots of the diagonal elements of $\mathbf{A}^{-1}$ are then reported as the errors in the fit parameters.

Estimate dY Error Method
With the simplifying assumption that $\sigma_{y_i} = \sigma_y$ for all y_i, then the matrices **A** and **B** become:

$$\mathbf{A} = \begin{pmatrix} m & \sum x_i & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \cdots & \sum x_i^{m+m} \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^m y_i \end{pmatrix}.$$

Once the fit parameters have been determined, common error σ_y are estimated by calculating $\sigma_y^2 = \frac{1}{N-m+1}\left\{\sum_i \left(y_i - a_o - a_1 x_i - a_2 x_i^2 - a_3 x_i^3 - \ldots - a_m x^{m-1}\right)^2\right\}$ for an m^th^ order polynomial fit to a set of N data points. This can be put back into matrix **A** and inverted to determine errors in the fit parameters.

X and Y Error Bars Error Method
This has not been implemented for polynomial fits.

References

J.H. Williamson, "Least-Squares fitting of a straight line," Can. J. Phys. **46**, 1845-1847 (1968).

P.R.Bevington and D.K. Robinson, Data Reduction and Error Analysis For the Physical Sciences. Boston: McGraw-Hill, 1992. pp.96-109.