Linear regression

> This problem sheet is based on [1]. See any standard statistics textbook for discussions of this problem.

You have a set of $N > 2$ points (x_i, y_i), with known Gaussian uncertainties σ_{yi} in the y direction, and no uncertainty at all (that is, perfect knowledge) in the x direction. You want to find the function $f(x)$ of the form

$$f(x) = m\,x + b \quad , \tag{1}$$

where m is the slope and b is the intercept, that *best fits* the points. The following is the standard practice.

Construct the matrices

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} \quad , \tag{2}$$

$$\boldsymbol{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \cdots \\ 1 & x_N \end{bmatrix} \quad , \tag{3}$$

$$\boldsymbol{C} = \begin{bmatrix} \sigma_{y1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{y2}^2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_{yN}^2 \end{bmatrix} \quad , \tag{4}$$

The best-fit values for the parameters m and b are just the components of a column vector $\boldsymbol{X}$ found by

$$\begin{bmatrix} \mathrm{b} \\ \mathrm{m} \end{bmatrix} = \boldsymbol{X} = \left[\boldsymbol{A}^\top \boldsymbol{C}^{-1} \boldsymbol{A}\right]^{-1} \left[\boldsymbol{A}^\top \boldsymbol{C}^{-1} \boldsymbol{Y}\right] \quad . \tag{5}$$

This procedure is not arbitrary; it minimizes an objective function χ^2 ("chi-squared"), which is the total squared error, scaled by the uncertainties

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i)]^2}{\sigma_{yi}^2} \equiv [\boldsymbol{Y} - \boldsymbol{A}\,\boldsymbol{X}]^\top \boldsymbol{C}^{-1} \, [\boldsymbol{Y} - \boldsymbol{A}\,\boldsymbol{X}] \quad , \tag{6}$$

that is, equation (5) yields the values for m and b that minimize χ^2. This, of course, is only one possible meaning of the phrase “best fit”.

When the uncertainties are Gaussian and their variances σ_{yi} are correctly estimated, the matrix $\left[\boldsymbol{A}^\top \boldsymbol{C}^{-1} \boldsymbol{A}\right]^{-1}$ that appears in equation (5) is just the covariance matrix (Gaussian uncertainty variances on the diagonal, covariances off the diagonal) for the parameters in $\boldsymbol{X}$.

Exercise 1: Using the standard linear algebra method of this Section, fit the straight line $y = m\,x + b$ to the x, y, and σ_y values for data points 5 through 20 in the file `straightline.dat` That is, ignore the first four data points, and also ignore the column for σ_x. Make a plot showing the points, their uncertainties, and the best-fit line. What is the standard uncertainty variance σ_m^2 on the slope of the line?

Exercise 2: Repeat Exercise 1 but for all the data points in the `straightline.dat` file. What is the standard uncertainty variance σ_m^2 on the slope of the line? Is there anything you don't like about the result? Is there anything different about the new points you have included beyond those used in Exercise 1?

Exercise 3: Generalize the method of this Section to fit a quadratic (second order) relationship. Add another column to matrix $\boldsymbol{A}$ containing the values x_i^2, and another element to vector $\boldsymbol{X}$ (call it q). Then re-do Exercise 1 but fitting for and plotting the best quadratic relationship

$$g(x) = q\,x^2 + m\,x + b \quad .$$

References

[1] David W. Hogg, Jo Bovy, and Dustin Lang. Data analysis recipes: Fitting a model to data. *ArXiv e-prints*, 1008:4686, August 2010.