A simple spreadsheet-based program for fitting Physics Labs data

J.-M. Frère, ULB
2006

Why?

I was frustrated to realize that most entry-level programs either don't take into account the error bars in the fitting of the data, or only do so for the "vertical" axis, not for the "horizontal". The issue is even more funny in the special case of a basic use of Excel, which allows both to include error bars, and to perform fits to a curve, - just that the fit does not take into account the error bars.

What

The simple program found in the spreadsheet (to be downloaded separately) gives an example of a 2-parameter fit: (assume we are fitting data corresponding to y=f(t), like a free-fall experiment, for instance: $y=f(t)=V0\ t + G/2\ t^2$).
It takes fully into account the error bars on the experimental points, both in y an t, using the effective variance method
(see: J Orear , Am J. Phys 50(10)oct 1982, p 912, D R Barker and L M Diana, Am J. Phys 42;p224, appendix 3 of http://www.astro.washington.edu/ivezic/Astr507/orear.pdf)
for a reference on using spreadsheets for non-linear fitting, see also
http://faculty.washington.edu/ambrown/

I have now (may 2009) checked that these worksheet could be used with the open-source program OpenOffice.Org , but it requires the non-linear solver. A preliminary version of this add-on is available at
http://extensions.services.openoffice.org/project/NLPSolver
(the horizontal error bars may not show up, but the optimization of the chi-squared works)

How (as in "how to"… more details later)

The raw data are introduced into the table, (the place is indicated), to replace the (fancy) data provided. (these data correspond to a fancy value of G, to be found at best on another planet, and are there just for the example).
If less data points are available, it is a simple matter to delete a corresponding number of row.
It on wants however to fit more data points, 2 steps are necessary :
-- insert the requested number of rows in the table
-- select the whole table (from the first line of data, excluding the titles) down to the last introduced row, and extend down the definitions of cell contents (CTRL-D, or EDIT -> Fill-> Down)
-- just for the readability of the graphs , the measurements should be introduced with increasing values of t . If is is not the case, a "sort" can fix that.

The proposed sheet then takes the (arbitrary, introduced by hand) values of V0 and G to generate a graph (seen on "chart1"), including the data points, their errors, and a curve corresponding to f(t). Of course, with the arbitrary values taken, the curve has no reason to match the data points.

Students can then experiment by trial and error to modify VO and G to match the curve at best.
While doing so, they are encouraged to watch the “chi squared” (named “sum of weighted squares in the spreadsheet) , which measures in some way (see below, for details) the departure of the curve from the data points.
Alternatively, they can use the “solver” (tools -> solver) to minimize automatically the chi squared. In principle, when selecting the cell in red, and calling the Solver, the options should be set right, but it is best to check that it is set to minimize this cell, while varying VO and G). At this moment, the adjustment of the curve is performed.

The program then provides the “best” values of V0 and G, and a suitable plot in “cahrt1”.
It may be interesting for the students to see how much departure from these values is allowed, as a function of their variables.

This is performed in an approximate way (see details below), and “chart 2 “ provides the intervals of confidence in the form of an ellipse in the plane (V0,G) , corresponding approximately to an increase in the chi squared of 1 unit.

Quality of the fit

Once the students are explained the meaning of the chi squared (see details below), they can be taught that a typically expected value for a good fit would be a chi squared per degree of freedom around 1.
Much higher values hint at inconsistency of the data with the functional form, while much smaller values may suggest an overestimation of the error bars (for instance, the stated error on may “meters” involves both an overall normalisation – with no statistical effect- and a much smaller departure from linearity).

Sensitivity to individual measurements,

It can be part of the exercise to see how adding either very precise or very unprecise measurements affects (or does not affect) the fit.
For this, we have included an additional spreadsheet, where we constructed an example showing how 2 “bad” points with large errors could completely fool the usual automatic fitting feature of Excel , while having little impact on the present construction.
Similar checks can be made about the relevance of the errors on absissae.

Relevance to beginner’s lab data

It may be argued that in “normal” circumstances, the non-linearity appearing in the present fit is of little consequence (and often, the automatic fit can give a close answer). As is easily seen form the info below, the non-linearity is here very minimal indeed, and appears only through the weighting of the points. It would disappear for a fit to a straight line and equal error bars.

Situations where the non-linearity of the data or the difference in error bars are however easily met.
For instance, we can consider the period of a pendulum as a function of the angle :
$T = T_0 (1+ \Theta^2/16 + ..)$.
If the fit is done as a function of the angle Θ measured , the errors on Q will generally be equal in absolute value (say $\Delta\Theta = 0.02$ radians), but the fit will be to a parabola. To get closer to a "pencil and paper" approach, one would be tempted to use a graph (T vs $z= \Theta^2$), in which case we get a straight line, but errors depending strongly on z : $\Delta z = \Delta Q^2 = 2\ \Theta\ \Delta\Theta$.

Some details about the spreadsheet,

Fitting another function

Is easy enough, but in this simplistic spreadsheed requires to change the definition in the function and of its derivative, which appear in columns E and F respectively, but also (for the calculation of the confidence interval) in I,J,K .

What is fitted.

In principle, one should minimize the distance of the measured point to the curve, both in y and in t, weighted by the respective absolute errors (assumed to be 1 σ) . This would however involve inverting the curve y=f(t) , a process which is trivial for a straight line, still feasible (with an ambiguity) for a 2^{nd} degree polynomial, but represents a non-linear problem of its own for more general curves.
For this reason, the method of "effective variance" is used here, and the error in t is "converted" to an error in y : how much the shift in t would affect the error in y.
The two errors in y (one resulting from the positioning of a photocell, the other from the time measurement for constant position of the cell) are then added quadratically, as independent errors.
Thus, the effective variance used is :
$\sigma_{eff}^2 = \sigma_y^2 + (df/dt * \sigma_t)^2$

The quantity to minimize (effective χ squared) is then

$$X = \sum ((y_i - f(t_i))/\sigma_{eff\ i})^2$$

Confidence interval ellipse
In principle, one should give the confidence interval (and the correlation) for the fitted parameters (here, VO and G).
I found it easier for the student to provide the curve corresponding to a given increase in the χ squared. (here, 1)
It would be quite easy to do this again using the SOLVER, and asking to adjust the valued searched to exactly the previously found minimum +1, and to repeat the operation in various directions in the (V0,G) plane (determined by a parameterκ)
For simplicity again, we have here instead approached the c squared function around its minimum by a paraboloïd. For this, we used small departures from the minimum to estimate the second derivatives : d^2/dy^2 , d^2/dt^2, $d^2/dy\ dt$, and then built a rough picture of the ellipse by scanning the plane in directions governed by a parameter κ.
The result is provided in chart 2