Least-Squares Fitting of a Straight Line

Experiments typically require the establishment of some relationship between physical parameters – for example, the period of a pendulum and its length. All of the experiments in the Physics 2B laboratory require the fitting of a straight line to experimental data, from which the physical quantity of interest, and its uncertainty, can be determined.

But given a series of linear data points, each with an uncertainty, how does one determine the "best" straight line through the set of data? The best fitting straight line is clearly that which can be drawn passing through, or close to, all the data points. But:

- no two people will draw the same "best" line through a given data set
- if the uncertainty on each data point is different, how do we take this into account in determining the "best" straight line?
- finding the uncertainties in the gradient and *y*-axis intercept of the graph is cumbersome, and tends to overestimate their values

To avoid these problems we can give a mathematical statement of what the "best fitting" straight line means, and then use a computer to determine it for us.

Chi-Squared: Figure 1 shows an *x-y* graph with a line passing close to the data points. For each individual value of *x*, labeled x_i, there are two values of *y*: y_{io} refers to the *observed* value, that is, the one measured in the experiment, and y_{ic} refers to the *calculated* value of *y* found using the equation of a straight line

$$y_{ic} = mx_i + c$$

The difference between the observed and calculated *y*-value of each data point is called the *residual* and is given by

$$\Delta y_i = y_{io} - y_{ic}$$

We clearly want of straight line to have small values of Δy_i for each data point. And we state that the *best fitting* straight line will be the one where the sum of the squares of the residuals (squared so that each has a positive contribution to the sum) is smallest. If each data point has an uncertainty in its *y*-coordinate of σ_i, then we can define a quantity χ^2 (chi-squared) as:

$$\chi^2 = \sum_{i=1}^{i=N}\left[\frac{(y_{io} - y_{ic})}{\sigma_i}\right]^2 = \sum_{i=1}^{i=N}\left[\frac{(y_{io} - (mx_i + c)}{\sigma_i}\right]^2$$

Where the sum is over the *N* data points.

Key Point: The best fitting straight line is the one which gives the minimum value of χ^2.

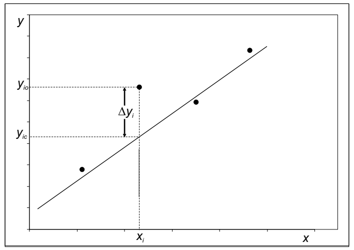

Figure 1: *x-y* graph showing the residual Δy_i.

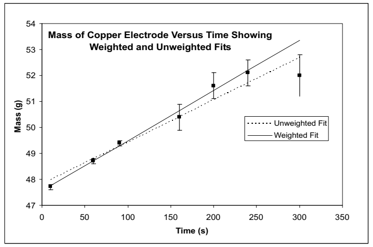

Figure 2:Least-squares fit to data showing both weighted and unweighted fits.

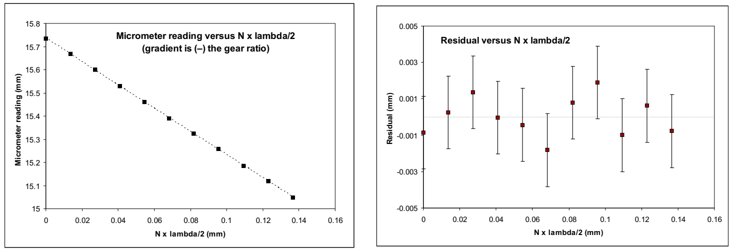

Figure 3: (Left) Linear LS fit to data from the Michelson interferometer experiment. The fit is excellent, and a plot of the residuals (Right) reveals no systematic deviations, but suggests that the uncertainties on each data point have been slightly overestimated.

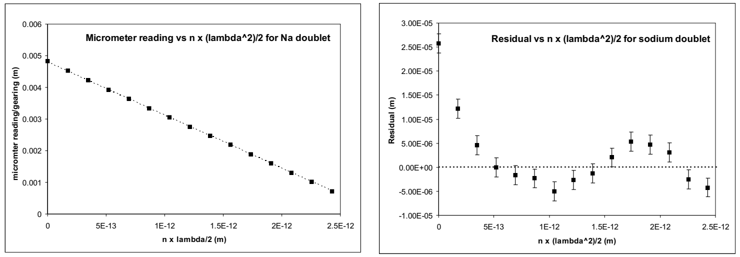

Figure 4: (Left) The linear LS fit to this data from the Michelson interferometer experiment looks excellent, but a plot of the residuals (right) from the fit shows a clear systematic deviation from a straight line.

Weighted and Unweighted Fits

In an *unweighted* LS fit we assume that the uncertainty on each data point is the same, and all data points are then treated equally. This is the type of fit performed by the standard **LINEST** function in Excel.

In a *weighted* LS fit, we assume that the uncertainties are not the same and weight each data point separately. Data points with small uncertainties are then fitted at the expense of those with larger uncertainties. This is the type of fit performed in Excel by the **LSFIT** function. This is a non-standard addition to Excel, and was developed in the Dept of Physics at the University of Texas,

Weighted and unweighted fits to the same experimental data are shown in Figure 2. Note that the weighted fit preferentially fits the data with the smaller uncertainties at the expense of the data point at *t*=300s which has the largest uncertainty.

The Expected Value of χ^2

For *N* data points fitted by a straight line having 2 adjustable parameters (*m* and *c*), the expected value of χ^2 is *N*–*2*. This is also known as the number of degrees of freedom. If χ^2 is larger or smaller than this then this indicates that something is wrong with the assumptions that you have made.

- If $\chi^2 > N-2$ then the model is not capable of representing the data to the accuracy suggested by the error bars. Thus, the error bars are *too small* for the scatter of the data.

- If $\chi^2 < N-2$, then the errors have been overestimated, i.e. the fit to the data is *too* good, and the error bars are *too large* relative to the scatter of the data.

The value of χ^2 can be estimated from a plot of the residuals after the least-squares fit.

- If $\chi^2 \approx N-2$, then ~2/3rds of the data should lie within ±1σ of the zero.
- If $\chi^2 > N-2$ then <2/3rds of the data lie within ± 1σ of the zero.
- If $\chi^2 < N-2$ then >2/3rds of the data lie within ± 1σ of the zero.

Figure 3 shows a linear fit to data obtained from a Michelson interferometer experiment. The fit to the data is excellent, and a plot of the residuals shows that *all* the data points are within ±1σ of the best-fitting straight line, suggesting that the uncertainties in the micrometer readings of ±0.002mm have been overestimated. We would therefore expect the value of χ^2 to be smaller than 11–2=9. Calculation gives a value of χ^2= 3.1.

Figure 4 shows a linear fit to further data obtained from the same experiment. Although the linear fit to the data *looks* excellent, a plot of the residuals shows a serious systematic misfit. We would therefore expect χ^2 to be greater than 13–2=11. Calculation gives χ^2= 244.

This final example demonstrates the invaluable information that is *only* visible in a plot of the residuals.