Least Square Fit to a line.

Independent variable x. Given a specific independent value x_i, the corresponding "y_i" value is given by the formula $(\text{line})_i = a + bx_i$. For a set of N points, specified by positions x_i, y_i (the real data), some of or none of the points may fall on a "best" line. We assume that the values of the independent variable are well known, and if there is uncertainty in these we relate it to a corresponding uncertainty in the dependent variable.

How do we select parameters a,b to give a line that fits best to the data, and what does best fit mean? It is up to us to decide and define.

We consider how far the measured value y_i (which occurs at x_i) is from the predicted "theoretical" value of our line $(\text{line})_i$.

$$\Delta y_i = y_i - \text{line}_i = y_i - (a + bx_i)$$

It is easy to envision a horizontal line through a set of data, with half the data points above and half below the line. If we do that (horizontal line) then the Δy_i values would cancel since some are positive and others negative, but this certainly is not a "best line". There are many other "bad" lines that can be drawn with the sum of the Δy_i adding to zero. Instead consider looking at $(\Delta y_i)^2$ and adding these for the set of data. So we shall sum the squares of the "residuals" (vertical distance of each point from a hypothetical line). Sometimes this is called "chi squared".

$$\chi^2 = \sum_i (\Delta y_i)^2 = \sum_i (y_i - a - bx_i)^2$$

The right hand side is often (more rigorously) divided by σ_i^2. Here σ_i is the standard deviation for each point. I have not included this point by point standard deviation so my "chi squared" is not properly normalized. That's OK, because we usually approximate the standard deviation as equal for each of the data points. Since we are going to minimize the "chi squared" (take derivative and set to zero) anyway, the constant standard deviation would cancel out when we set the derivative to zero.

We must minimize the "chi squared" with respect to each of the parameters a and b in order to find the minimum "chi squared". In other words, we want to find the choice of a and b gives the line that reduces the sum of the squares of the residuals.

$$\frac{\partial(\chi^2)}{\partial a} = 0 \text{ leads to } \sum_i (y_i - a - bx_i) = 0$$ **1. SHOW**

and

$$\frac{\partial(\chi^2)}{\partial b} = 0 \text{ leads to } \sum_i x_i (y_i - a - bx_i) = 0$$ **2. SHOW**

Rearranging 1 yields:

$$\sum_i y_i = \sum_i a + \sum_i bx_i = aN + b\sum_i x_i$$

and 2 yields

$$\sum_i x_i y_i = \sum_i a x_i + \sum_i b x_i^2 = a\sum_i x_i + b\sum_i x_i^2$$

The last two equations are simply two simultaneous equations for a and b. The summations are messy but simply constants that relate to the specific data set. Solving for a and b gives.

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

3. SHOW FOR a, b, Δ (NEATLY,

CAREFULLY, PLENTY OF PAPER AS NEEDED)

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

$$\Delta = N\sum_i x_i^2 - \left(\sum_i x_i\right)^2$$

And now without statistical justification or derivation, I will give formulas used to calculate the square of the standard deviation in each of the parameters a and b.

$$\sigma^2 = \frac{1}{N-2}\sum_i (y_i - a - bx_i)^2$$

and

$$\sigma_a^2 = \frac{\sigma^2}{\Delta}\sum_i x_i^2$$

and

$$\sigma_b^2 = N\frac{\sigma^2}{\Delta}$$

That's it. Note that in the calculations for a and b and uncertainties, there are only five types of sums that need to be evaluated (five columns on a spreadsheet).

Problems: **steps 1,2,3 above (write it up neatly)**
and 4 below.

4. Use the speed of light data from our first homework to generate the summation each of the quantities needed to find a, b, and Δ, and the uncertainties in a and b. That is generate the five columns with values of x_i, with x_i^2, y_i, $x_i y_i$ and $(y_i$-a-$bx_i)^2$. Sum these columns, placing the resulting sum at the bottom of each. You will print out the spreadsheet with numbers and each column labeled clearly.

Now determine a, b, and the standard deviations for each by hand using the results of your spreadsheet. Compare to results you had using the “fit” button previously---they should be very very close.