MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Physics Department

Physics 8.01T Fall Term 2004

Introductory Error Analysis

Linear Regression: Least-Squares Fits

It is common to try to fit known curves to a set of data points $\{(x_i, y_i)\}_{i=1}^{n}$. Indeed, you have already done this using *DataStudio* several times.

If the desired fit is to y as a linear funcion of x, the "best" fit, that which minimized the total error (as described below) is $y = a + b\,x$, where

$$a = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i\,y_i\right)}{n\,\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}, \qquad b = \frac{n\,\left(\sum x_i\,y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{n\,\left(\sum x_i^2\right) - \left(\sum x_i\right)^2},$$

where all sums are from $i = 1$ to n. These expressions are often written more compactly as

$$a = \frac{\langle y\rangle\langle x^2\rangle - \langle x\rangle\langle xy\rangle}{\langle x^2\rangle - \langle x\rangle^2}, \qquad b = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\langle x^2\rangle - \langle x\rangle^2},$$

where if g is any function of x and y, $\langle g\rangle = \frac{1}{n}\sum g_i$.

These calculations are simple but tedious, which makes them good candidates for either a calculator or computer. Any scientific calculator will have a linear regression feature (but possibly with different names for the slope and intercept). As we have seen with *DataStudio*, fitting a "best" straight line is quick and easy.

If *DataStudio* is not available, any spreadsheet program should be able to do the calculations for you. What follows is a short guide to using Xess, the default spreadsheet program on server at MIT.

When the window appears, start entering data. It's recommended that you enter the values of x_i in one column and the values of y_i in an adjacent column. If you have exported your data to a "tab-separated value" file (one with a `.tsv` extension), you can import your data directly into your Xess session; go to "File" on the menu bar.

If, for instance for Experiment 4, Circular motion, your data consists of the measures values of r_m and the calculated values of ω, you can manipulate the data to find ΔX and F. If your values for r_m are in Column B, starting with cell B1, click on cell C1 and in the edit line enter

`=B1-4.1`

followed by a return (my unstretched spring length was 4.1 cm). Click again in cell C1 and copy the formula, either from "Edit" on the menu bar or Control-f. Then drag into the rest of the cells in Column C where you want the formual applied. What you see is the value of the cell immediately to the left minus 4.1. The terminology is that the "`B1`" is a "relative" attribute.

Similarly, click on cell D1 and enter

`=C1*8.6*A1*A1/1000`

followed by a return. (This assumes that the values for ω are in Column A.) I used a mass of 8.6 gm, and the division by 1000 will give a force in newtons.

To plot the calculated force as a function of the extension, highlight Columns C and D, go to "Graph" on the menu, select "New Graph" and "Scatter Graph" to get the plot. Adjust the features as desired using "Edit" and "Options" on the graph window.

To find the best-fit line, click in an empty cell and enter

`=@LINCOEF(C1..C5,D1..D5)`

to display the slope (a) and intercept (b). (I had four non-zero measurements, so I needed to give the range in the columns to include all five points).

To graph this line, go to another empty cell and enter

`=@LINFIT(C1..C5,D1..D5)`

to calculate $a\,x_i + b$ for each value in Column C. Of course, you could use the best-fit parameters to do this yourself, but since it's a built-in function, we'll use it.

To graph this line on the same graph as the calculated force *vs* displacement graph, go to "Edit" "Data Sets" on the graph window. At the top, you'll see a slide bar under "1". Either move the slide to the right or click to the right of the bar to bring up Data Set 2. In the "X Data" field enter `C1..C5` and in "Y Data" enter `F1..F5` (or the corresponding cells where you put the `LINFIT` values). For the best display, select "Line" for "Segment Type" and "None" for "Marker."

In *Data Studio* you may have noticed that when you make a linear fit to data, the window describing the fit has an entry for "r", the *correlation coefficient*. Of course Xess can do this. In an empty cell enter `=@CORR(C1..C5,D1..D5)`. The calculation of the correlation coefficient is a bit messy, but the spreadsheet doesn't mind at all.

Determining the Best-Fit parameters

Suppose you have plotted n data points (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), and you suspect a linear relation between the values of x and y; how can you determine what linear relation best fits your data?

The most enlightening way is to plot the data and fit a line by eye. This line is characterized by a slope and an intercept that can be read off of the graph.

A more rigorous way is to define what we mean by the "best" straight line. Suppose a candidate line is given by $y = a + b\,x$. Then, each x_i gives a value for y. This will in general *not* be y_i (if it were, then *all* of the data would fit a line *exactly*), so denote the difference between y_i and its predicted value by $\epsilon_i = y_i - (a + bx_i)$ Here, ϵ_i is the "error in y_i." Then, consider the sum of the squares of these errors (we need to square the errors before adding since a negative error is as bad as a positive one, and we don't want them to cancel). This sum, denoted by Φ, is the quantity we wish to minimize; we want to find the values of a and b that give us the *Least Squares*

So, for a trial a and b, and a given set of n data points, we have

$$S = \sum_{i=1}^{n} \left(y_i - (a + b\,x_i)\right)^2 .$$

To make things easier, we'll introduce some compact notation. For any function g of x and y, let

$$\frac{1}{n}\sum_{i=1}^{n} g_i \equiv \frac{1}{n}\sum_{i=1}^{n} g(x_i,\, y_i) \equiv \langle g \rangle.$$

In this notation, then,

$$\frac{\Phi}{n} = \langle y^2 \rangle + a^2 + b^2\langle x^2 \rangle + 2\,a\,b\,\langle x \rangle - 2\,a\,\langle y \rangle - 2\,b\,\langle xy \rangle.$$

(Note that $\langle x \rangle \langle y \rangle \neq \langle xy \rangle$; the product of a sum is not the sum of the products, but

you knew that.) Some rearrangement and a bit of razzle-dazzle yields

$$\begin{aligned}\frac{\Phi}{n} &= \langle y^2\rangle - \langle y\rangle^2 \\ &\quad + a^2 + 2\,a\,(b\langle x\rangle - \langle y\rangle) + (b\langle x\rangle - \langle y\rangle)^2 \\ &\quad + b^2\left(\langle x^2\rangle - \langle x\rangle^2\right) + 2\,b\,\left(\langle x\rangle\langle y\rangle - \langle xy\rangle\right) + \frac{\left(\langle x\rangle\langle y\rangle - \langle xy\rangle\right)^2}{\langle x^2\rangle - \langle x\rangle^2} \\ &\quad - \frac{\left(\langle x\rangle\langle y\rangle - \langle xy\rangle\right)^2}{\langle x^2\rangle - \langle x\rangle^2} \\ &= \langle y^2\rangle - \langle y\rangle^2 - \frac{\left(\langle x\rangle\langle y\rangle - \langle xy\rangle\right)^2}{\langle x^2\rangle - \langle x\rangle^2} \\ &\quad + \left(a + (b\langle x\rangle - \langle y\rangle)\right)^2 + \left(b\sqrt{\langle x^2\rangle - \langle x\rangle^2} + \frac{\langle x\rangle\langle y\rangle - \langle xy\rangle}{\sqrt{\langle x^2\rangle - \langle x\rangle^2}}\right)^2 .\end{aligned}$$

This is a fancier version of completing the square, as we used in minimizing V_c to find the arithmetic mean as the "Best" average of a sample. Clearly, much hindsight was used to obtain the result in this form. What was done was to isolate the terms containing a, completing the square with those terms by adding $(b\langle x\rangle - \langle y\rangle)^2$, then completing the square with the b-terms.

The error will be minimized when both terms in parentheses are as small as possible in absolute value; that is, when they vanish. The "best" line is then that for which

$$b = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\langle x^2\rangle - \langle x\rangle^2} \qquad a = \langle y\rangle - b\langle x\rangle = \frac{\langle y\rangle\langle x^2\rangle - \langle x\rangle\langle xy\rangle}{\langle x^2\rangle - \langle x\rangle^2}.$$

The same result can be obtained from calculus by setting $\frac{\partial\Phi}{\partial a}$ and $\frac{\partial\Phi}{\partial b}$ to zero and solving the two linear equations for a and b.

Many calculators have this capacity, making determination of a and b wonderfully simple. If you're not sure how to do such things on your calculator, let us know (but *please* bring the instruction booklet).

The best way of all is to plot your points *and* use linear regression. This way, you can see if there are any points that are clearly wrong (but *never* discard a data point without good reason), or if the data even suggest a linear fit at all. Also, with an eyeball fit you can obtain an estimate of the error in a or b by seeing what range of a and b correspond to the extreme "reasonable" (as opposed to "best fit") lines to your data.