prepared by Annette D. Shine, August 2006

Fitting Experimental Data to Straight Lines (Including Error Analysis)

The purpose of this document is to assist students with statistical analysis of experimental data by listing some equations for straight line data fitting and error analysis. Personally, I find statistics texts very hard to plow through, so I am writing this document in understandable English for an engineer who wants to use statistics without being dazzled by the brilliance of the subject. If you find any mistakes, or want to suggest something else that needs to be included, or know of a book that is comprehensible and thus supercedes this treatment, please contact me.

Linear Least Squares Fitting of a Straight Line with Slope and Intercept

Any least squares curve- or line-fitting algorithm optimizes the constants of a fitting equation by minimizing the sum of the squares of the deviations of the actual (data) values from the values predicted by the equation. You probably know how to do linear least squares fitting of a straight line already, since most scientific calculators and graphing software packages do this automatically for you. Nevertheless, I will present it here so that: (1) you will be aware of assumptions inherent in use of the canned programs, (2) you can verify with your own calculations that you get the same answers as the canned programs, and (3) I can build on this base for cases where some of the "canned" assumptions are not valid.

Given: A set of n experimental data points,

x_1, y_1
x_2, y_2
$\vdots \quad \vdots$
$\vdots \quad \vdots$
x_n, y_n

where x is the independent variable (i.e., the thing you fix or consider fixed, such as time when you are measuring reaction kinetics, or voltage when you are using a pressure transducer), and y is the dependent variable (i.e., the thing that you want to determine, such as extent of reaction or pressure). The x_i values are assumed to be listed from lowest to highest. (It is not really necessary here to list the points in order of increasing x_i, but it will be in a later part of this document.)

Furthermore, let's assume that the relationship between x and y is a linear one (if it's not, fitting a line to the points is worthless).

Let $y = ax + b$

be the equation of the best fit line to the data. We wish to determine the values of both the slope a and the intercept b. If we assume that each data point carries equal weight, i.e., each y_i point has exactly the same actual (not relative) error associated with it, then we find a and b by minimizing the sum of the squares of the deviations of the actual values of y_i from the line's calculated value of y. The formulas for a and b are:

$$a = SLOPE = \frac{(n\sum x_i y_i) - (\sum x_i)(\sum y_i)}{(n\sum x_i^{\,2}) - (\sum x_i)^2} \qquad (1)$$

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

In all equations, the summation sign is assumed to be from i=1 to i=n.

For example, consider this actual calibration data for the vortex flowmeter from the frictional losses experiment in junior lab:

Voltage (x_i)	Flow rate, liter/s (y_i)
1.01	0.00
1.27	0.19
1.85	0.58
2.38	0.96
2.83	1.26
3.13	1.47
3.96	2.07
4.91	2.75

The linear least squares fit to this data gives the line:

$y_i = 0.70303729738\ x_i - 0.7153519908$ (liters/sec)

(It is always a good idea to carry along as many significant figures as possible during statistical calculations because truncation errors may be significant when subtracting two nearly equal values. I've included all these figures above so you may check the calculation yourself, if you wish.)

If you use the linear curve-fitting routine in Excel, you get:

$y_i = 0.703\ x_i - 0.7154$ (liters/sec)

with a correlation coefficient $R^2 = 0.9999$.

Uncertainties (Errors) In Calculated Slope and Intercept

Suppose that the calculated slope and/or intercept from the "canned" equations above was really the experimental quantity of interest, say a reaction rate constant or an initial reaction rate. In this case, you will want to determine the error associated with the slope or intercept so you can present the experimental uncertainty, i.e., to give a plus-or-minus value. The formulas (for points with equal error, as above) are:

$$ERR - a = SLOPE \cdot ERROR = S * \sqrt{\frac{n}{(n\sum x_i^{\,2}) - (\sum x_i)^2}} \qquad (3)$$

$$ERR - b = INTERCEPT \cdot ERROR = S * \sqrt{\frac{\sum x_i^2}{(n\sum x_i^2) - (\sum x_i)^2}} \qquad (4)$$

where

$$S = \sqrt{\frac{\sum (y_i - ax_i - b)^2}{n-2}} \qquad (5)$$

Note that **S** is the square root of the quantity found by dividing the sum of the squares of the deviations from the best fit line, by the number of data points you have *beyond* the minimum required (two points determine a straight line) to fit the specified curve. The quantities a and b are those calculated for the best fit line.

For the above data, **S** = 0.011769957 liter/sec (Note that **S** has the units of y). The associated errors in the slope and intercept are

slope error = 0.003343664 liter/sec-Volt and

intercept error = 0.009842206 liter/sec.

This means that the relative errors are

relative slope error = 0.003343664/0.70303729738 = 0.48% and

relative intercept error = 0.009842206/0.7153519908 = 1.38%.

Note that the slope has a smaller relative error than the intercept, so that you can get more reliable estimates if you plot the data in such a fashion that the quantity you want to extract is the slope. Alternately, be sure you always do error analysis when using a calculated intercept value, since the error can be large, especially if the fit is not too good. (The above example has excellent data fit, as seen by the R^2 value.)

Linear Least Squares Fitting and Error of a Straight Line Which MUST Go Through the Origin.

This is the same case treated above, except that now we FORCE the line to go through the point (0, 0). Because we specify that the intercept is 0, the only parameter we can determine is the slope a, i.e., we find the best "a" value for the equation

y = ax.

For example, in a chemical reaction, we know that the rate of disappearance of species A must be zero if the concentration of A is zero. In the past, you may have just used the canned fitting routines, hoping that the intercept value came out small anyhow; the following are the correct equations to use:

$$a = SLOPE = \frac{\sum x_i y_i}{\sum x_i^{\,2}} \tag{6}$$

$$ERR - a = SLOPE \cdot ERROR = S_{0,0} * \frac{\sqrt{\sum x_i^{\,2}}}{\sum x_i^{\,2}} \tag{7}$$

Here, $S_{0,0}$ has the same meaning as **S** in eq (5) above, except that only one additional point is needed to draw a straight line through the origin, so

$$S_{0,0} = \sqrt{\frac{\sum (y_i - ax_i)^2}{n-1}} \tag{8}$$

Uncertainties Resulting from Interpolation and Extrapolation of Straight Line Data

Since many measurements are now made with electronic instruments, engineers are frequently required to assess the reliability of an indirectly measured variable that was arrived at through comparison with an instrument calibration curve. Alternately, it is sometimes necessary to extrapolate experimental data to arrive at a prediction of a variable at some condition beyond the experimentally measured range. When the experimental or calibration data can be fit to straight lines, the formula for estimating error at a point is the same whether the point is interpolated or extrapolated . Specifically,

Given: A set of n experimental data points, rank ordered so that x_1 is the lowest x_i value, and x_n is the highest x_i value.

Let y = ax + b

be the best fit straight line to the data, with a and b calculated as above. If you wanted to predict a value for y, say y*, given a specific value of x, say x*, you would obviously predict that the most probable value of y* when x=x* is

$y^* = ax^* + b.$

This would be true regardless of whether x* fell within the experimental range of x_i. Now, suppose you want to know how "correct" this y* value is, i.e., you want to know how much uncertainty is associated with the predicted y* value.

For a confidence interval of 100(1-α)% (e.g., at the 95% confidence level, α=0.05), the ± uncertainty associated with y* is:

$$y^*\_ERR = t_{\alpha/2,n-2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{n(x^* - \bar{x})^2}{n\sum x_i^{\,2} - (\sum x_i)^2}} \qquad (9)$$

Where S is given by eq (5) if the line was NOT forced through the origin, $\bar{x}$ is the average x-value of the experimental data points, i.e.,

$$x = \frac{\sum x_i}{n}$$

and

$t_{\alpha/2,\ n-2}$ is the critical value for the t distribution, which can be looked up in statistical tables (e.g., see http://socr.stat.ucla.edu/Applets.dir/T-table.html). The value of n-2 is called the number of degrees of freedom (since these are the "extra" points beyond what was needed to determine the original straight line), and is often given the symbol ν or df in statistical tables. Be careful in using t-distribution tables that you use the correct numerical value for α/2 (i.e., HALF the difference from 100% of the confidence interval).

If the line fit to the original data was forced through the origin, then the number of degrees of freedom is only n-1, and S is calculated from eq (8).

Consider again the voltage/flow rate vortex flowmeter calibration data given above. Suppose now that in the laboratory, you measured three different flow rates with the vortex flowmeter, such that the voltage readings were:

1.50 V, 3.45 V and 4.61 V. What would be the expected flow rate, together with its +/- error term for each voltage?

Suppose we want a 95% confidence interval; then $\alpha = 0.05$. From the calibration data above, n = 8, $\Sigma x_i = 21.34$, $\Sigma x_i^2 = 69.3154$, $\Sigma(y_i - ax_i - b)^2 = 0.000831191$, S = 0.011769957 liter/sec (Note that S has the units of y), and $t_{0.025,\ 6} = 2.447$, so the three predicted values of flow rates, together with their estimated uncertainties due to the calibration curve fitting are:

x* (V)	y* (liter/s)	+/- y* (liter/s)	Relative error in y (%)
1.50	0.3391	0.013962	3.7 %
3.45	1.710138	0.012436	0.73 %
4.61	2.5256644	0.018876	0.75 %

You can also use the t-equation above to find uncertainties in extrapolated values of y, provided you know that the linear relationship holds in the extrapolated regime (e.g., you couldn't extrapolate friction factor vs. Reynolds number data taken for 1<Re<2000 out to Re=10,000, since data were taken in the laminar range and the extrapolation goes to the turbulent range). In fact, we cannot reliably extrapolate the flowmeter data above, because if we go to x* lower than the experimental range, the flow would be predicted to be negative, which is not physically realistic, while if we go to higher voltages, we exceed the 5-Volt limit of the instrument.

If you care to do so, you can use the t-equation to reconstruct the equation above for error in the intercept of an extrapolated line. The error prediction is one standard deviation, which corresponds to a 68% confidence interval, and effectively assumes an infinite number of data points; in this case $t_{0.16,\ \infty} = 1$.

Reference for this section:
Probability and Statistics for Engineering and the Sciences, Second Edition, by Jay L. Devore, Brooks/Cole Publishing Company, p. 478 Monterey, CA 1987 (ISBN 0-534-06828-6)

Weighted Least Squares Straight Line Fitting

All the above equations assumed that each data point had the same amount of absolute (not relative) error associated with it. This is rarely the case in practice. It is much more common for *relative* errors to be similar, or for errors to be larger at the extreme ends of the measurement range. In any case, it makes sense in curve fitting to give the least amount of weight to points that are the least reliable. This is properly accomplished statistically by weighting each point by the inverse square of its standard error when calculating the best-fit slope or intercept.

To use this method, you must first establish the standard error of *each* point; this quantity will be called e_i. If you made repeat measurements, you should use the standard deviation. If the points you are plotting are actually derived quantities (e.g., the slopes of previously plotted best-fit lines), then you should use error propagation to get e_i (e.g., use the slope error above). All previous equations assumed e_i was a fixed constant, so it didn't really matter whether the (fixed) error resided with x_i or with y_i; here, the error is presumed to reside with y_i. Hence, for curve-fitting purposes, enter your data points so that y_i is the coordinate with the most error. (This is usually, but not always, the case "naturally," since we tend to use simple things such as time or reciprocal temperature as the independent x coordinate.) Note that e_i is an absolute, not relative, error, so it has the same units as y_i.

For weighted linear regression, the best fit values of the slope a and the intercept b are then given by:

$$a = SLOPE = \frac{(\sum \frac{x_i}{e_i^2})(\sum \frac{y_i}{e_i^2}) - (\sum \frac{x_i y_i}{e_i^2})(\sum \frac{1}{e_i^2})}{(\sum \frac{x_i}{e_i^2})^2 - (\sum \frac{x_i^2}{e_i^2})(\sum \frac{1}{e_i^2})} \qquad (10)$$

and

$$b = INTERCEPT = \frac{(\sum \frac{x_i y_i}{e_i^{\,2}}) - a(\sum \frac{x_i^{\,2}}{e_i^{\,2}})}{\sum \frac{x_i}{e_i^{\,2}}} \tag{11}$$

or alternately

$$b = INTERCEPT = \frac{(\sum \frac{x_i}{e_i^{\,2}})(\sum \frac{x_i y_i}{e_i^{\,2}}) - (\sum \frac{y_i}{e_i^{\,2}})(\sum \frac{x_i^{\,2}}{e_i^{\,2}})}{(\sum \frac{x_i}{e_i^{\,2}})^2 - (\sum \frac{x_i^{\,2}}{e_i^{\,2}})(\sum \frac{1}{e_i^{\,2}})} \tag{12}$$

The uncertainties in the slope a and intercept b of a least squares line that was weighted by the individual errors of the points are given by:

$$aERR = slope\_error\_of\_weighted\_line = \sqrt{\frac{\sum \frac{1}{e_i^{\,2}}}{(\sum \frac{x_i^{\,2}}{e_i^{\,2}})(\sum \frac{1}{e_i^{\,2}}) - (\sum \frac{x_i}{e_i^{\,2}})^2}} \tag{13}$$

$$bERR = \text{int}ercept\_error\_of\_weighted\_line = \sqrt{\frac{\sum \frac{x_i^{\,2}}{e_i^{\,2}}}{(\sum \frac{x_i^{\,2}}{e_i^{\,2}})(\sum \frac{1}{e_i^{\,2}}) - (\sum \frac{x_i}{e_i^{\,2}})^2}} \tag{14}$$

If the line MUST pass through the origin, then the slope is determined from:

.

$$a = slope\_of\_line\_through\_origin = \frac{\sum \frac{x_i y_i}{e_i^2}}{\sum \frac{x_i^2}{e_i^2}} \tag{15}$$

The error in the slope of a weighted least squares line passing through the origin is given by:

$$err\_a = slope\_error\_of\_weighted\_origin\_line = \frac{S_{0,0}}{\sqrt{\sum \frac{x_i^2}{e_i^2}}} \qquad (16)$$

The parameter $S_{0,0}$ is defined in Eq. (8).

Note that, since all of the above relationships, including the error terms, involve only a few different sums, they should be relatively easy to program on a spreadsheet.