Fitting Functions to Data

Introduction:

Scientific experimentation involves the measurement of the value of a property, while changing other experimental parameters in a systematic fashion. For example, in chromatography, the quantity HETP (H) is measured for different carrier gas linear velocities (u) in an experiment to find the minimum value of H keeping all other chromatographic parameters constant.
The relationship between H and u conforms to a function of the form;

$$H = A + \frac{B}{u} + Cu$$

Here H (dependant variable) is said to be a function of u (independent variable). The relationship between H and u is not simple and direct.

To take another example, the variation of the resistance, R, of metals with temperature T is as;

$$R = a + bT$$

It is a direct linear relationship. Experiments are carried out to verify these relationships and determine the coefficients such as A, B, C (and a, b); by following values of H with a varying values of u (and R with varying T in the latter case). The first step in the analysis of a set of data obtained from experiments is to make a plot of dependant vs. the independent variable. Data points constitutes pairs of data values; (H',u'), (H'',u''),.. etc., and ideally such plots should yield smoothly varying curves or lines. It is common knowledge that experimental data does not necessarily conform to such smooth plots.

In a typical experimental investigation, a value for the dependent variable (say H) is found for each value of the independent variable (here, u). A single measurement is hardly sufficient to get an accurate value for the quantity H corresponding to a certain u. A single measurement invariably carries an indeterminate error. This error may be positive, negative or zero. The errors in H values will result in the distribution (scatter) of data points on either side of the ideal value, which is expected to lie on the smooth curve. The deviations of data points from the expected ideal values are usually randomly distributed. In this course we will consider only situations where the associated errors possess this type of random character.

The objective of an experiment is to use the experimental data points (which inevitably carry errors) to arrive at the assumed form of the function of interest, i.e. to determine the coefficients such as A, B, C, etc. With data inherently carrying errors it is not possible to determine values for the desired coefficients with no errors. The practical approach is to find the best possible values for the coefficients (and their associated errors) that will generate a function to follow the observed data points in the closest possible manner, at a certain confidence level. This process of finding the best fitting curve, which is the determination of the said coefficients, is known as regression analysis.

Regression analysis is a mathematical technique (see text for details) that determines the values of coefficients (constants) of the function. The coefficients determined are such that the best fit functions produces the minimum sum of squares of deviations for the data set. The assumption here is that the standard error associated with the dependant variable values are random and follow a Gaussian distribution pattern and the standard deviations of all data points are the same. If the errors associated for all the data points are not statistically the same, the weighted least square technique must be used for regression analysis.

Before subjecting a data set to regression analysis it is imperative to reject outliers from the data set. Direct observation data plots and Q test is of great help to recognize the outliers.

Regression Calculations from Spreadsheet Programs:

The built in functions of spreadsheets offer an easy way to perform the regression analysis. In Microsoft Excel for example, after entering the data in the spreadsheet in an orderly fashion (in columns) they are processed with the built in macros/commands. The output will yield the coefficients of the ‘best fit’ equation and the respective errors. Use the coefficients from the regression to generate a continuous best fit curve. Superimpose the experimental data points (as points) on the best fit plot (curve/line), document all spreadsheets, worksheets etc.

Exercises:

The following data set (y vs. x) follows a straight line (linear) relationship; y = ax + b. Determine the best fit line and the associated errors of the coefficients using a spreadsheet program (Excel) or a symbolic math program, (e.g. Mathcad).

Data Set 1

x	4.1	5.0	6.0	7.2	8.1	9.0	9.9	10.8	12.0.0	13.0	14.0
y	5.4	5.8	6.02	6.3	6.5	7.1	7.5	7.9	8.0	8.6	9.0

Spreadsheet Approach:

- Enter the data in columns; x and y in columns A and B respectively; label columns
- Plot a x-y scatter graph (points)
- Name the graph, labels
- Inspect the preceding graph, reject outliers
- Carry out the regression after rejecting the outliers
- Regression output will generate values for coefficient a, b (as the constant) and their errors, these error values may be used in the calculation in the last part of this exercise.
- Enter the formula for the best fit calculation in the column C and generate the corresponding best fit points.
- Superimpose the best fit (xy graph, continuous) graph
- Generate a hard copy of the plot and report the results.

(When reporting experimental data give all observations and make a note of the outliers).

2. The data set 2 given below fits an equation of the form;

$$A = A_f \frac{1}{1-\frac{A_0 - A_f}{A_0}\exp\left(\frac{-k_{obs}A_f t}{\varepsilon}\right)}$$

The two parameters to be determined via the best fitting procedure are A_f and k_{obs}. Given time (independent variable) and corresponding A value (dependent variable), use the 'Microsoft Excel Solver' to determine the best fitting values. Assume A_0 = 0.50723 and ε = 1020.

Solver, is an Excel add-in, and is already installed on all of the computers… however; it is not enabled by default. Enabling the solver add-in is as follows,

1. Click the "File" tab and choose **Options**
2. Choose **Add-Ins**, Select *Solver Add-in*
3. Click the "Go" button at the bottom of the Add-ins window
4. Check **OK**

Data Set 2

Time/s	0	60	120	180	240	300	360	420
Absorbance, *A*	0.50723	0.48962	0.4733	0.45877	0.44608	0.43346	0.42337	0.4129

Time/s	480	540	600	660	720	780	840	900
Absorbance, *A*	0.40375	0.39626	0.38828	0.38049	0.37381	0.36777	0.36227	0.3568

Time/s	960	1020	1080	1140	1200
Absorbance, *A*	0.35213	0.34729	0.34312	0.33925	0.33557

- Enter x and y values in columns A and B starting from the second row. Label the rows.
- Enter labels A0, epsilon, Af and kobs in cells G2, G3, G5, G6 respectively
- Enter values forA0, epsilon in cells H2 and H3, (actual values/constants); Af and kobs in cells **H5 andH6** (initial guess for parameters to be determined) respectively
- Assign "Excel names" to values of A0, epsilon, Af and kobs; e.g. Click on H2, *Formulas*, → *Define Name*, type A0, → *OK*.
- Name x variable as vector of name t; Highlight A2:A22, *Formulas*, → *Define Name*, →OK, type t, *Add*, *OK*
- Calculate the A values for initial guesses of parameters; Click C2, type in cell C2, =Af/(1-((A0-Af)/A0*EXP(-kobs*t*Af/epsilon))) ; the function to which the data is being fitted.
- To find the best fit values for Af and kobs (optimization) set up an 'indicator' of the error between the calculated values assuming the function (formula) and the experimental values; In cell D2 type the formula (B2-C2)^2 and copy the formula up to D22.
- In cell D23 sum up D2 through D23, this will be the measure of the error.

- Click on D23, *Data* → *Solver* (in Analysis Tab).
- Solver Parameters windows pops up. Set as the target the cell $D$23.
- In the same window select, By Changing Cells **$H$5; $H$6**
- Click *Options*; Set Min 0.000001 (a low value, ideally it is zero!)
- Click Min radio button, Click *Solve*.
- In Solver Results window select Keep Solver Solution and Answer, Click *OK*
- A Report would be generated as a worksheet.

Plot the data points and the best fit curve.

3. The following data set fits an equation (a polynomial in x) of type y = a + b/x +cx.

Data Set 3

x	40.1	50.1	75.1	90.23	103.2	128.2	152.3	162.6	200	225	250	295.6	359.3
y	6.12	5.2	4.1	3.27	3.25	2.95	2.94	3.01	2.7	2.65	2.8	3.11	3.23

The regression of this data set with a spreadsheet is similar to previous exercise. There are three coefficients, namely a, b and c associated with variables, x^0, x^{-1} and x^1. Given x (independent variable) and corresponding y (dependent variable) is 'unique'. The Excel add-on Solver can be used to determine the best values (of the best fit curve).

4. The data file provided (*Data Set 4*) as an Excel file: Fityk data.xls is the raw series of x, y data output from a HPLC run.

- *Assuming* that each peak approximates sufficiently well to a Gaussian function, determine the peak value and the FWHM (2×HWHM) of the peaks using the ***Fityk*** program. (Under some conditions the baseline corrected x,y data is fitted, however. It may not be for the data in this exercise).
- The actual chromatographic peaks are *asymmetric*. Determine the areas of the asymmetric peaks.

Fityk is program for nonlinear fitting of analytical (especially peak-shaped) functions to data. It can also be used to remove the baseline from data. It fits functions of various forms such as Gaussian, Lorentzian, and Voigt etc. Fityk also supports user-defined functions. (http://fityk.nieto.pl/)

next page, please

Fityk Documentation

Enter the (x,y) data set in column A (x) and B (y) in an Excel spreadsheet. Save the spreadsheet. Save the spreadsheet again in the .csv format. Fityk accepts Excel .csv files as an input.
Invoke the Fityk program.

- *Data* → *Load File* (select the xxx**.csv** file)
  Open in new slot → *Close*
- Click [*auto-add*] button as many times as appropriate. Refrain from being trigger happy, however. Note the addition of a Gaussian function (default) every time it's clicked.
- Click [*start fitting*] button once. Note the Gaussian function characteristics starting with selecting the 'functions' tab. Clicking of the red squares would result in the appearance of the fitted Gaussian parameters of that function in the window below. To further manually change the Gaussian function parameters edit values in the fields at the right-bottom window.
- Print the deconvoluted output plots and the overall fit; *Session* → *Page setup* (select *black lines with white background*); *Session* → *Print* (select the printer).
- Save results: *Functions* →*Export Peak Parameters* (type in the filename; xxx**.txt**)
- Save session: **Session → Save session** (enter a session name)