

Fitting Data to a Mathematical Function

Fall 2009

Introduction

In many of the laboratory exercises in this course you will be collecting data representing a phenomenon in the real world, and attempting to find the mathematical function that gives the best approximation to those data. A simple version of this procedure occurs when you graph linear data and find the slope and intercept of the line that best approximates those points. You know from our first lab, however, that it is not always easy to determine exactly which line best describes your data. Today we will look at this problem in more detail, and learn how to use a computer to simplify this task.

Procedure

Data Set #1: A Linear Function

Consider the experiment of measuring the position x of an object as a function of time as it moves along the x -axis. If the object starts at $t = 0$ with constant x -velocity, we expect the data to be modeled by the equation $x = v_o t$, where v_o is the adjustable parameter. If v_o is adjusted properly, there will be as many data points above the line you have drawn as there are below, and the sum of the differences between data points and corresponding points on the line (the *residuals*) will be very close to zero. The sum of the square of the residuals (SSR – also called *Chi Square*, χ^2) can be thought of as a function of the parameter v_o and the goal in obtaining the best-fit line is to minimize the SSR.

1. *Graphical Analysis:* Using the values in Data Set #1 (found on the *Fitting Data* page), and assuming the values labeled t_i represent time in seconds and the values labeled d_i represent distance (in meters), plot the points on a graph and draw your best guess for the line fitting these data points (forcing your best-fit line through the origin will better match your later results). Calculate the slope from your graph. This will be your best value of v_o from a hand-drawn slope.
2. *Analysis by Calculation:* We must first derive the relationships we will need. Using calculus, the best value of v_o is calculated by minimizing the SSR. We define the SSR to be the square of the distance between each of the 11 points and the line $x = v_o t$, added up:

$$\text{SSR} = (x_1 - v_o t_1)^2 + (x_2 - v_o t_2)^2 + \dots \quad \text{Eqn. 1}$$

To get the best v_o , differentiate SSR with respect to v_o and set it equal to zero:

$$\begin{aligned} d(\text{SSR})/dv_o &= 2(x_1 - v_o t_1)(-t_1) + 2(x_2 - v_o t_2)(-t_2) + \dots = 0 \\ &= 2\{-(t_1 x_1 + t_2 x_2) + v_o(t_1^2 + t_2^2 + \dots)\} = 0 \end{aligned} \quad \text{(factor out } -2 \text{ and clean up)}$$

What is inside the parentheses must equal zero if SSR is a minimum:

$$v_o(t_1^2 + t_2^2 + \dots) - (t_1 x_1 + t_2 x_2) = 0$$

Or, solving for v_o gives:

$$v_o = \frac{\sum_i t_i x_i}{\sum_i t_i^2} \quad \text{Eqn. 2}$$

An essential portion of our analysis will involve the calculation of the *uncertainty* in the slope, which for Data Set #1 is accomplished by calculating the standard error in v_o :

$$\boxed{\text{Standard Error} = \sqrt{\frac{SSR}{(N-1) \sum t_i^2}}} \quad \text{Eqn. 3}$$

Where N is the number of data points you have ($N = 11$ for Data Set #1).

Now for the calculations: Use Eqn. 2 to calculate v_o from the data (you may find it useful to use a table for your calculations). Then use Eqn. 1 to calculate the SSR using your *calculated* value of v_o (**from Eqn. 2 – don't use v_o from your graph**). Finally, use Eqn. 3 to calculate the uncertainty in v_o . *As will be explained later, you will state the value of the slope +/- twice the standard error in v_o .* Use the memory function of your calculator so that you don't introduce rounding errors during your calculations.

3. *Computer Analysis:* Now you will perform the same calculations using the computer. We will be using the program KaleidaGraph to analyze data in this lab. Follow *all* of the instructions in *Graphing & Curve Analysis Using KaleidaGraph* to do the following for Data Set #1:
 - Create a spreadsheet from the data.
 - Create a graph.
 - Determine the best-fit function, and find the *fit parameters* from this function.
 - The *fit parameters* are the coefficients that are adjusted to minimize the function, giving you the lowest SSR value. In future experiments, you will calculate experimental values such as position, velocity, and acceleration from these parameters.
 - Calculate the uncertainty in the slope.
 - Calculate the residuals and the SSR of the fit.
 - Print the graph.

KaleidaGraph is going through essentially the same process you did to calculate the best-fit slope for the line, but you will find the computer does it more quickly.

4. In your report, record the *three* values you determined for v_o from steps 1, 2 and 3, and record the *two* values you calculated for the SSR and slope uncertainty from steps 2 and 3.

Data Sets #2 and #3: Non-Linear Functions

Sometimes it is difficult to determine what function will best approximate a set of data simply by looking at it. Graphing the data can give you some idea, but if the data do not fall exactly along a line (and they rarely do in the real world), it can be difficult to tell the difference between power, exponential, and polynomial functions.

5. Use KaleidaGraph to plot Data Sets #2 and #3 on *separate* graphs, then try different functions to determine which one gives you the best fit for each set. Be sure to try *all* the fits to determine the one that appears best by visual inspection. Calculate the residuals and SSR for the best of each set. *Don't bother printing these graphs;* in your report, record the type of fit you chose, the best-fit equation, the values of the fit parameters and the SSR.

Note: If you wish to save time, you will find two files in T:\Phys 151 that contain Data Sets # 2 and #3!

Discussion

- Your discussion should consist of a summary of the results you obtained for each data set:
 - *Data Set #1:* The hand drawn graph, the KaleidaGraph plot, v_o (three values), SSR (two values), the uncertainty in the slope (two values). *Don't forget the units!*
 - *Data Set #2 & #3:* Record the fit equation (with numerical values) and SSR for each data set.