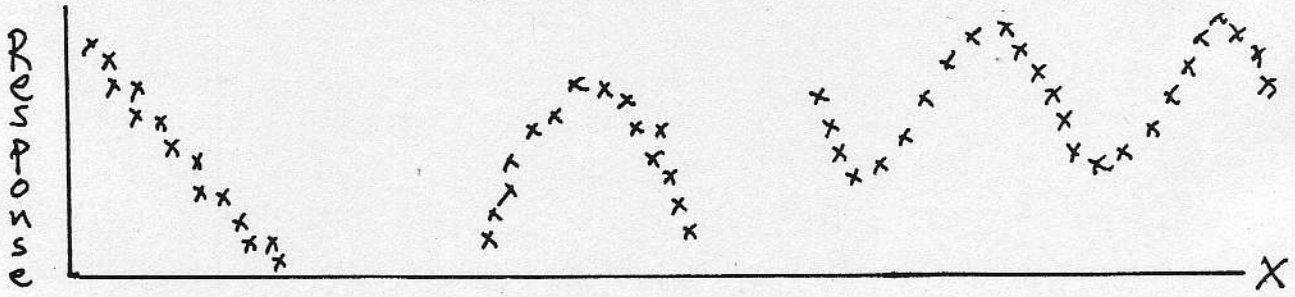
OAS Lecture 5: Modeling Data

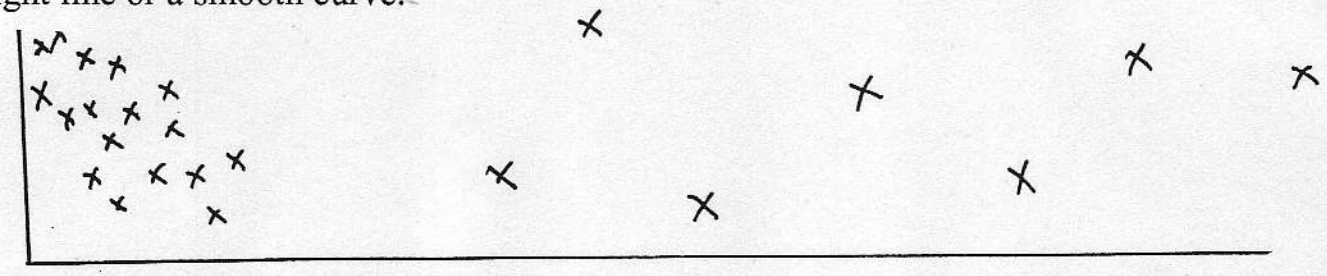
You may have noticed that much of what has been discussed to this point in the course has to do with designing and performing an experiment. As a result of forming hypotheses and conceiving of well-thought-out experiment designs is that you are able to produce lots of data.



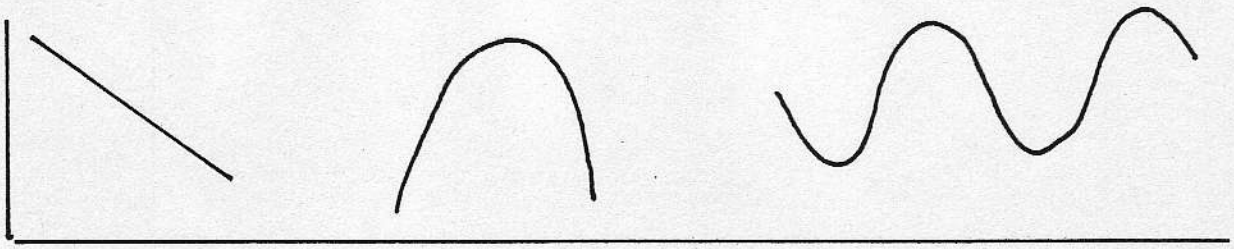
Shown above are the kinds of plots of data that you might generate during this course. Notice I haven't assigned specific labels to the axes. This is because whatever kind of data you collect, it tends to mimic simple mathematical functions. Functions like straight lines, or parabolas or sine waves or hyperbolas. It is in this observation that the beautiful interplay of science and mathematics is seen. Imagine, all those elementary functions you learned about in math class are recreated when you collect data in the laboratory.

$y = mx + b$ $y = kx^2$ $y = A\sin(\omega t)$ $y = k/x$

Of course a big complication is that the data you acquire in the lab isn't in the form of a nice straight line or a smooth curve.



Instead it is a collection of data points, maybe too few data points, maybe with too much noise.



The question becomes, can I figure out exactly what the elementary function is that best models my data?

In answering this question it is important to remember that when we predict a model function using deductive reasoning, we already know that the data should be a straight line or a parabola or a sine wave. Thus we know that energy is related to the square of velocity:

$$E = 0.5mv^2$$

The relationship between pressure and temperature is a straight line function of the form

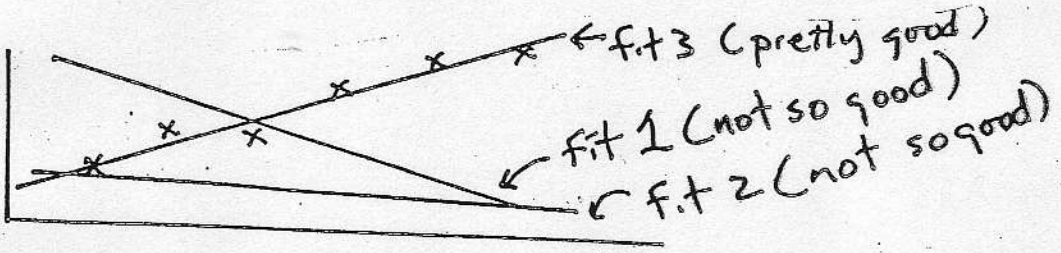
$$P = kT$$

We even know that the oscillations of a pendulum are modeled by a damped oscillator:

$$y = A\exp(t/\tau)\sin(\omega\tau + \phi)$$

Notice that for every one of these functions, have a collection constants (m, k, A, τ, ω, ϕ) that scale the magnitude of the response. Mathematicians are less concerned about what these constants are and instead refer to them as **coefficients** and typically assign them generic names like: a_0, a_1, a2, a_3, …., a_n. It is the value of these coefficients that were are after when we try to fit curves to data. Thus, when we know that a function is a straight line, we can then ask the question, what is the slope (a_1) and what is the y-intercept (a_0). Remember in earlier days you might have referred to the slope and y-intercept as m and b as in $y = mx + b$.

The point is that there is one combination of coefficients that best describes the data we collect. In the example below, notice that of the three straight lines that are drawn, only one seems to fit well to the data that is collected. Only one set of a_1 and a_0 fits the data.



So how do we identify the best values for the coefficients? One obvious way is to fit the line we draw as closely as we can to the data points, as we did for one of the examples above. This is know as "chi (χ) by eye" for reasons you will understand in a few minutes.

Is there a better way to fit the data? A more systematic and accurate way? The answer is yes. The mathematical way that experimentalists most commonly use to fit functions to data, to curve fit, to determine the best coefficients, is called "**least squares**" . The rest of this lecture is devoted to understanding how least squares is performed. You will see that it isn't very hard at all to implement least squares. Interestingly, you will also notice that it is an "**optimization**" technique, because, after all it is finding the best, or optimum, function to fit the data. There is an optimization technique used to perform least squares. We could use a simplex, but that isn't the best approach to use when you already have a function. Remember that when we have a function, the best approach is the derivative method which is the most efficient of the optimization tools. It allows the best function to be fit in a single experiment. Think back to

our optimization discussion--brute force took 1000s of experiments, simplex took a couple dozen, derivative methods take one experiment.

With this introduction I'm going to start using more sophisticated language as I develop least squares, but remember, all I am doing is coming up with a way to find the best coefficients for a function used to fit experimental data.

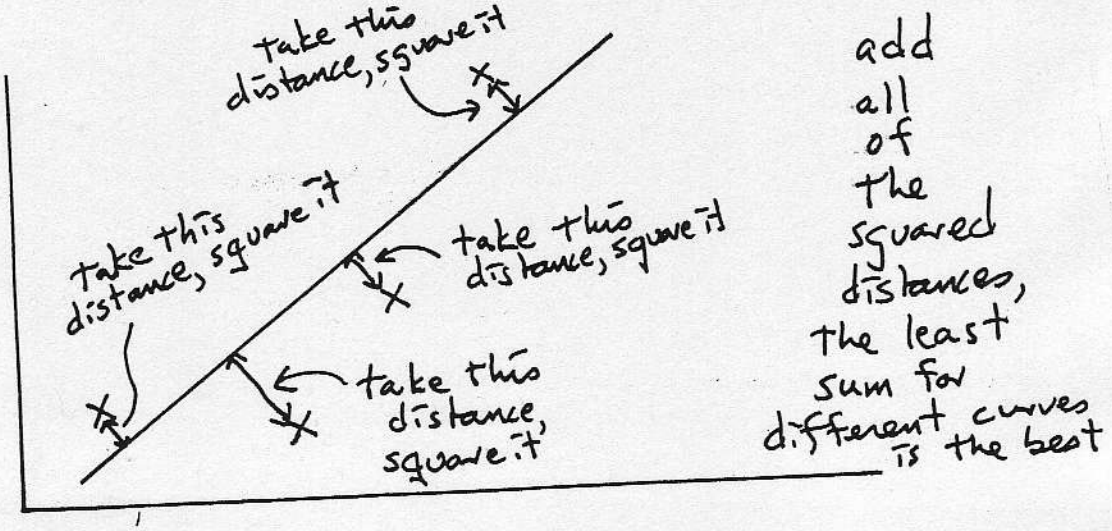
Fitting linear functions. Often it is desired to condense or summarize a group of observations in a model with adjustable parameters (coefficients). You can either be fitting data with an arbitrary, but convenient function such as a polynomial, or the fit may be made to a function related to underlying theory. Whatever you do, the basic approach is to choose a figure of MERIT FUNCTION that measures agreement between experimental observations and the model. Typically, a small merit function value indicates a good fit. In other words, a kind of minimization or optimization is performed to obtain a best fit. As we will see, LINEAR LEAST SQUARES methods provide a single pass determination of the minimization (optimum fit), while
non-linear methods are distinguished by the need to apply an iterative determination of the minimum. (We will save our look at non-linear methods until later in this life, or likely, another life far, far away.) Instead we concentrate in this lecture on LINEAR processes. Recognize that linear does not mean the function is a straight line, but rather linear least squares refers to a linear combination of any functions. For example:

$$y(x) = a_1 + a_2 x + a_3 x^2 + \ldots\ldots a_m x^{m-1}$$

is an example of a linear function with a_k coefficients. Or it could be a function that combines sines and cosines. It is only the LINEAR dependence on coefficients a_k that defines the ability to use the LINEAR LEAST SQUARE method.

Now because the date you collect has uncertainty, there will be some measurement (noise) error that will require an assessment of the correctness of the model. In other words, we will need a statistical justification of a goodness of fit. This is often overlooked and instead often people stop at a determination of the best coefficients (a_ks). this is known as "χ (chi) by eye". Unfortunately, given the lack of time available, we too will succumb and develop linear least squares without doing statistics. But know that they are out there in fully developed analyses.

The Least Squares Method. The approach we will introduce for modeling the use of least squares as merit function to estimate goodness of fit. (Can you think of other possible



merit functions?) For a fit of N data points (x_I, y_I) to a model with M adjustable parameters a_k, there is a functional relationship.

$$y(x) = y(x; a_1 \ldots\ldots a_m) \qquad \text{eq1}$$

with a least squares fit

$$\Sigma[y_I - y(x; a_1 \ldots\ldots\ldots a_m)]^2 \qquad \text{eq2}$$

Now we are going to want to find the coefficients, a_k, that will give you the smallest value for eq 2. But wait!! What is the one thing that you actually remember from calculus? *If you want to find the minimum of a function, you take the derivative and set it equal to zero.* Well what we want to do is find the minimum of eq 2. So just take the derivative of eq 1, set it equal to zero, and there you have it, the best fit values for the coefficients.

The great thing about this is that you don't just apply this approach to straight line functions like $y=a_2x+a_1$. You can apply it to any linear combination of functions, and as long as you can take a derivative of the square of the functions in eq. 2 you can solve for the best fit of that data!!

Thus for example, take any function with a_k coefficients you want to find for a best fit,

$$y(x)= y(x; a_1, a_2, \ldots\ldots\ldots a_m) \qquad \text{eq3}$$

Examples:

$y = a_2x + a_1$
$y = a_3x^2 + a_2x + a_1$
$y = a_2e^{-ax} + a_1 x^2$

You want to find a minimum to the merit function as in eq 4.

$$\chi^2 = \sum \left[\frac{y_i - y(x_i; a_1, a_2 \cdots a_n)}{\sigma_i} \right]$$

(often times σ is not known so let σ =1). Now take the derivative and set to zero in eq 5.

$$0 = \sum_{i=1}^{n} \left(\frac{y_i - y(x_i)}{\sigma_i^2} \right) \left(\frac{\partial y(x_i; a_1, a_2 \cdots a_n)}{\partial a_i} \right)$$

What is left to do is to solve eq 5 and you have your single pass best fit of any combination of linear functions.

Now those of you who would rather forget any calculus you ever learned may not have a clue what I'm talking about right now, but trust me, this is amazing--if you have ANY collection of linear functions, this is a way to get the best possible fit in one iteration. (Actually the really amazing thing is that Gauss, the guy who thought up statistics of random variable at age 13, also thought all this up before he was 14.)

Example. Least Squares fit of a Straight Line. (Actually this kind of fit is so popular that it is given a special name, linear regression, and can be found on most everyone's calculator.)

Start with the straight line function, $y = a_2 x + a_1$. It has two coefficients, a_2 and a_1 that you want to find that give the best fit.

$$y(x) = y(x; a_1, a_2) = a_2x + a_1 \qquad \text{eq 6}$$

Now stick eq 6 into the merit function (eq 4) and you get eq 7

$$\chi^2(a_2, a_1) = \sum_{i=1}^{N} \left(\frac{y_i - a_2 x_i - a_1}{\sigma_i^2} \right) \qquad \text{eq 7}$$

And now taking derivatives with respect to a_1 and a_2 and setting to zero, we have two equation that need to be solved.

find best y-intercept $\Rightarrow 0 = \frac{\partial \chi^2}{\partial a_1} = -2 \sum_{i=1}^{N} \left(\frac{y_i - a_2 x_i - a_1}{\sigma_i^2} \right)$ eq. 8

find best slope $\Rightarrow 0 = \frac{\partial \chi^2}{\partial a_2} = -2 \sum_{i=1}^{N} \left[\frac{x_i (y_i - a_2 x_i - a_1)}{\sigma_i^2} \right]$ eq 9

It makes sense that because you have two unknown coefficients, you have two unknown equations. Fortunately, with two equations you can solve. Some substitutions will simplify eq 8 and eq 9:

let $S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$ $S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}$

$S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$ $S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}$ $S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$

And for those of you who fazed out till now, THE ANSWER:

let $D = (S)(S_{xx}) - (S_x)^2$

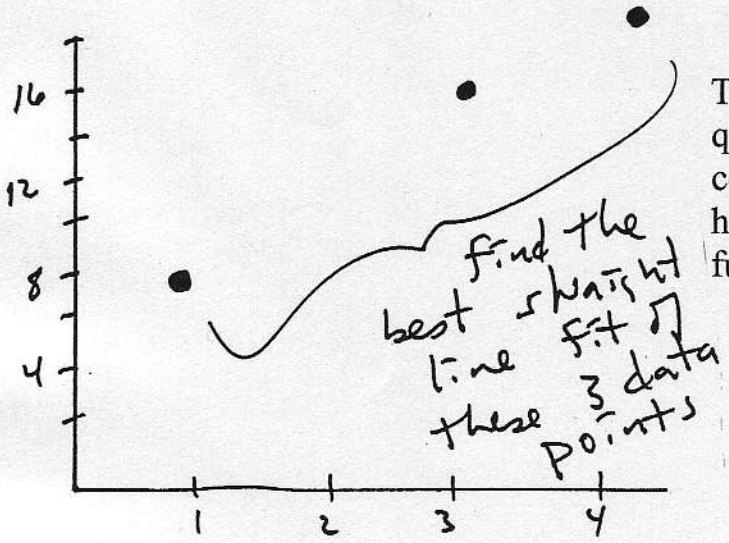
best coefficients for straight line

$a_1 = \frac{S_{xx} S_y - S_x S_{xy}}{D}$ y-intercept

$a_2 = \frac{S S_{xy} - S_x S_y}{D}$ slope

eq. 10

Now an Example: Find the best fit of the following three data points: (1,7); (3,15); (4,18).



The first thing we do is plot the data and ask the question, "what should I use for a model?" Of course the answer is, "a straight line" (since I haven't shown you how to fit any other kind of functions).

Okay, let's use equations 10 to solve. But first a guess.

What do you think the slope is? _____

What do you think the y-intercept is? _____

Now stick in the numbers:

First,

$S_x = 1 + 3 + 4 = 8$
$S_y = 7 + 15 + 18 = 40$
$S = 3$
$S_{xx} = 1 + 9 + 16 = 26$
$S_{xy} = 7 + 45 + 72 = 124$

Then $\Delta = (3)(26) - 64 = 14$

And

$a_1 = (26(40) - (8)(124) = 3.24$ (the y-intercept)

$a_2 = \frac{(3)(124) - (8)(40)}{14} = 3.71$ (the slope)

you can use any name you want for the coefficients,

$y = \underline{m}\,x + \underline{b}$ or $y = \underline{a} + \underline{b}x$ or $y = a_1 + a_2x$

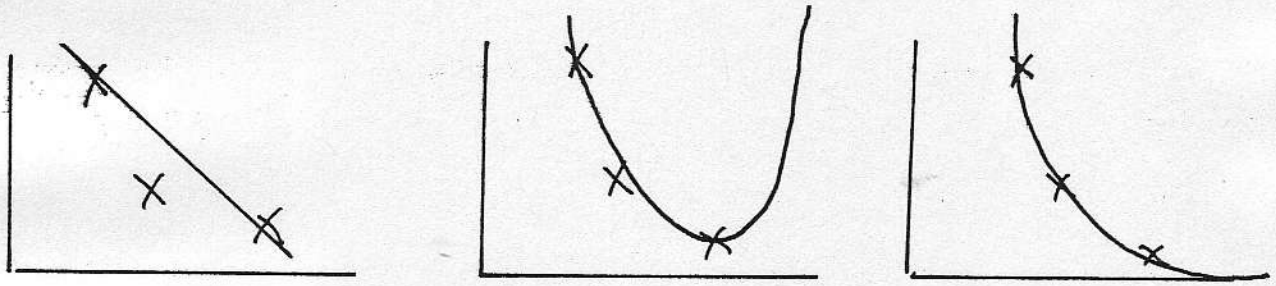
but

$y = 3.41 + 3.71x$

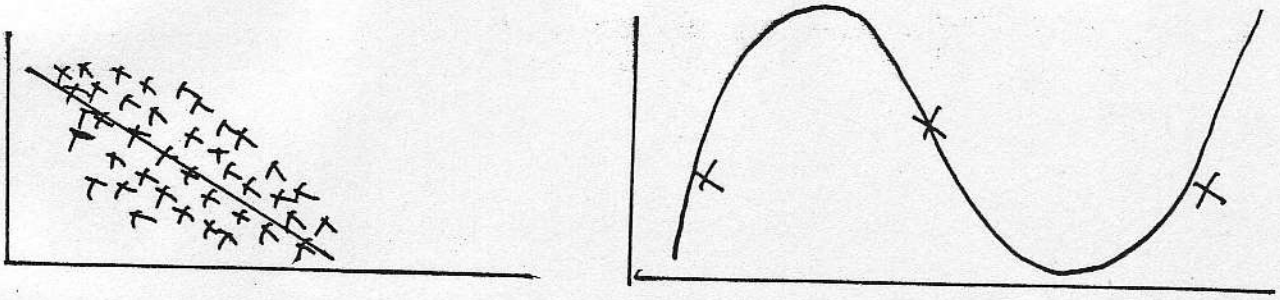
is the best fit of the data.

Extending least squares to more complicated functions. For some of you it will be enough that you have seen how easy it is to solve for a straight line function. But others will be excited to see that this general method can be applied to any function. All you need to do is to take the partial of equation 7 with respect to each coefficient (this means taking derivatives of simple functions which anyone can do) and then setting each partial equal to 0, as was done in equations 8 and 9 for two coefficients. Thus if you wanted the least squares solution for a parabola, $y = a_2x^2 + a_1x = a_0$, you would have three equations to solve for three unknowns, a_2, a_1, and a_0.

Using Mathematical Packages to Perform Least Squares. Most of you will probably see no reason to sweat through the details to curve fit some data points. It seems that every graphing or spreadsheet program these days has a sophisticated curve fitting program. However it is important to realize that the software can't do everything. In particular, you have to decide what the model function is. For example, for the data below, do you tell it to use a straight line, a parabola, an exponential function? The software will fit any data to any function--it is dumb as pitch. So it is up to you to know when and how to use the data.



Or consider what happens when you lack data points, or have a lot of noise in your data. You will find that the least squares will return answers to you, but that they may not be particularly valid, or may have a high degree of uncertainty.



In the end, curve fitting is a powerful tool, and you will use it a great deal on the computer in a few weeks. But for now, be aware that the most powerful software in the world can't tell you how to model your data, it cannot tell you what function to fit to the data, it cannot tell you how valid the fitted result are. It is up to you to take a reasoned approach to the use of powerful software packages that are increasingly available.