Concepts: Determining parameters in model, linear regression, details of fitting straight line to data.

Note: There are a few concepts here we haven't talked about in class yet, since most texts do not include a discussion of how regression works. I want you to have seen how regression works, but realize I will not test you on the details.

When using functions to model physical situations, you often have to determine some parameters in the model. The parameters are determined using some data that you know, typically in one of two situations:

- a small number of data points is provided which the model function must pass through, or
- a set of data point is provided which the model function must fit in some best manner.

In the latter case, the model function might not pass through any of the points in the data set.

Determining what the model function should be involves careful reading of the problem, or more typically knowledge of the subject area from which the problem arose (chemistry, biology, physics, psychology, economics, etc.). If you have a data set, you can do a scatter plot and use that to figure out a good place to start for a model function.

In this section, I want to focus on two situations–having the same number of data points as the number of parameters in the model (solve system of equations), and having more data points that the number of parameters (curve fitting–specifically, linear least squares regression).

Determining Parameters in a Model: System of Equations

If the number of points provided is the same as the number of parameters to determine, we use the points to determine a system of equations to solve to determine the parameters. The model function will pass through all the data points.

Example Find the quadratic function that passes through the points $(0, 0)$, $(1, 2)$, and $(2, -4)$.
The general quadratic function is $f(x) = ax^2 + bx + c$, with parameters a, b, c.
Use the data points to create a system of equations:
$(0, 0)$: $f(0) = a \cdot 0^2 + b \cdot 0 + c = 0 \to c = 0$.
$(1, 2)$: $f(1) = a \cdot 1^2 + b \cdot 1 + c = 2 \to a + b + c = 2$.
$(2, -4)$: $f(2) = a \cdot 2^2 + b \cdot 2 + c = -4 \to 4a + 2b + c = -4$.
Solving the system of equations yields $a = -4, b = 6, c = 0$.
Therefore, the function $f(x) = -4x^2 + 6x$ passes through the points $(0, 0)$, $(1, 2)$, and $(2, -4)$, which is easy to verify.

This process can be used for any type of model function, although the system of equations to solve can sometimes get complicated. For this to work, you need to have as many data points as unknown parameters, and the resulting model passes through all the data points.

Determining Parameters in a Model: Linear Regression Curve Fitting

If you are asked to fit a function to a set of data points, the process used is called *regression*.

If you are not told what the model function should be, can proceed by doing the following:

1. Make a scatter plot of the data.

2. Determine from the scatter plot what the function roughly looks like (linear, quadratic, sine, exponential, etc.).

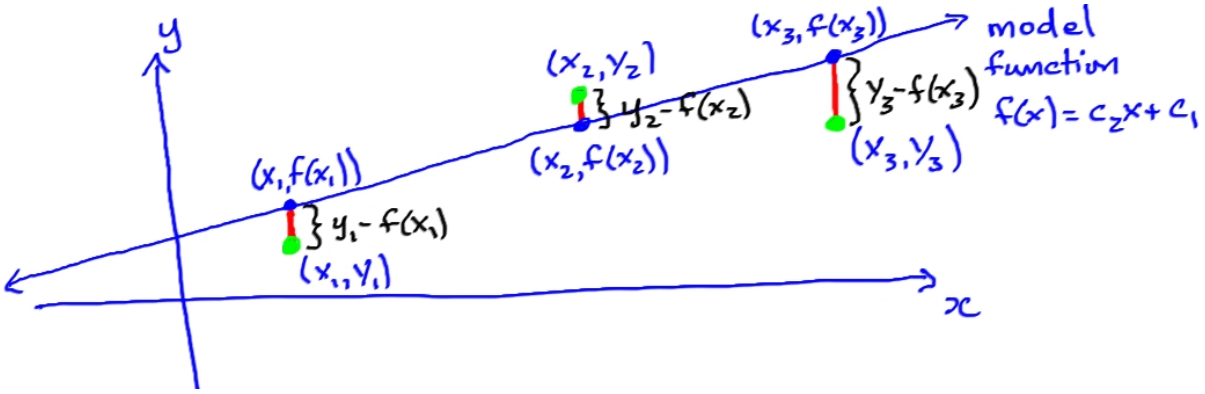
3. Transform a basic function of that type to best fit the data.

The last step is the hardest, and the one I want us to focus on for the case where the model function is a straight line, $y = f(x) = c_2x + c_1$, with parameters c_1, c_2.

Linear Least Squares Regression In General

If we are given n data points (x_i, y_i), $i = 1, 2, 3, ..., n$ and want to fit $y = f(x) = c_2x + c_1$, here is how we can proceed.

The diagrams shows three green data points (we will immediately generalize to n data points), the best fit line $f(x)$, and the vertical distance (in red) from the data points to the best fit line.



We choose to make the *best fit line* to be the one that makes the sum of all the vertical distances as small as possible (minimize):

$$E = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Note: We square the distances since we don't want to worry about $y_i - f(x_i)$ being positive or negative (ie., if the best fit line is above or below the data point). This is where the *least squares* in the name Linear Least Squares Regression comes from. In fact, we could use $E = \sum_{i=1}^{n} |y_i - f(x_i)|$ but then we would be doing least absolute value, not least squares!

Note: What makes it *Linear* Least Squares is the linear dependence on the parameters c_1, c_2. For example, model functions $f(x) = c_3x^2 + c_2x + c_1$ and $f(x) = ae^{3x} + b\sin(x)$ are also linear least squares, but $f(x) = ae^{bx}$ is nonlinear.

So we want to minimize E. You can use calculus to do this, or algebra. We will use algebra, although the calculus method is more appropriate for more complicated model functions. The key idea in the algebraic method is to

notice E is quadratic in c_1, and quadratic in c_2. The following manipulations show this.

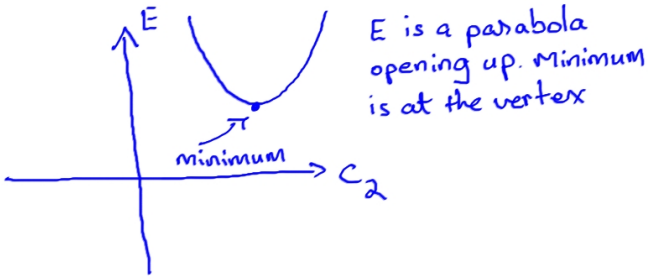
$$\begin{aligned} E &= \sum_{i=1}^{n}(y_i - f(x_i))^2 \\ &= \sum_{i=1}^{n}(y_i - c_2x_i - c_1)^2 \\ &= \sum_{i=1}^{n}(y_i^2 - 2c_2x_iy_i - 2c_1y_i + c_2^2x_i^2 + 2c_2c_1x_i + c_1^2) \\ &= \sum_{i=1}^{n} y_i^2 - 2c_2\sum_{i=1}^{n} x_iy_i - 2c_1\sum_{i=1}^{n} y_i + c_2^2\sum_{i=1}^{n} x_i^2 + 2c_2c_1\sum_{i=1}^{n} x_i + c_1^2 n \end{aligned}$$

In the last line we used the fact that $\sum_{i=1}^{n} 1 = n$.

Notice that E is a parabola in c_2:

$$E = c_2^2\left[\sum_{i=1}^{n} x_i^2\right] + c_2\left[2c_1\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_iy_i\right] + \left[\sum_{i=1}^{n} y_i^2 - 2c_1\sum_{i=1}^{n} y_i + c_1^2 n\right]$$

Since $\sum_{i=1}^{n} x_i^2 > 0$, E is a parabola opening up and looks something like the following as a function of c_2:



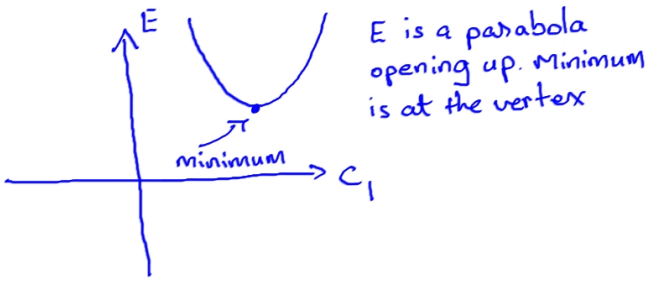
The minimum is at the vertex, and the vertex of $E = ac_2^2 + bc_2 + c$ occurs at $c_2 = -b/(2a)$ (something we will see later), so we have

$$c_2 = -\frac{b}{2a} = -\frac{2c_1\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_iy_i}{2\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_iy_i - c_1\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \tag{1}$$

Notice that E is also a parabola in c_1:

$$E = c_1^2[n] + c_1\left[+2c_2\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i\right] + \left[\sum_{i=1}^{n} y_i^2 - 2c_2\sum_{i=1}^{n} x_iy_i + c_2^2\sum_{i=1}^{n} x_i^2\right]$$

Since $n > 0$, E is a parabola opening up and looks something like the following as a function of c_1:



The minimum is at the vertex, and the vertex of $E = ac_1^2 + bc_1 + c$ occurs at $c_1 = -b/(2a)$, so we have

$$c_1 = -\frac{b}{2a} = -\frac{2c_2 \sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i}{2n} = \frac{\sum_{i=1}^{n} y_i - c_2 \sum_{i=1}^{n} x_i}{n} \tag{2}$$

Eqns. (1) and (2) form a system of two equations in the two unknowns c_2 and c_1. It is fairly straightforward to substitute Eqn. (2) into Eqn. (1) and solve for c_2. This gives the following equations for the parameters c_2 and c_1:

$$c_2 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$
$$c_1 = \frac{\sum_{i=1}^{n} y_i - c_2 \sum_{i=1}^{n} x_i}{n}$$

Summary: The linear least squares best fit function $f(x) = c_2 x + c_1$ to the data set (x_i, y_i), $i = 1, 2, 3, ..., n$ has

$$c_2 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$
$$c_1 = \frac{\sum_{i=1}^{n} y_i - c_2 \sum_{i=1}^{n} x_i}{n}$$

Determining how well this is as a fit involves more work than we will do in this class. However, I feel it is important that we understand how the fitting is done for this particular case, since it is used so often and only requires knowledge of parabolas to figure out!

Fitting other Model Functions

Other model functions can be fit to data, of course. Typically the calculus method is easier to implement to determine the equations of the parameters, but you should realize that is is still just minimizing the quantity E.

Eqns. (1) and (2) can rewritten as:

$$c_2 \left[\sum_{i=1}^{n} x_i^2\right] + c_1 \left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} x_i y_i$$
$$c_2 \left[\sum_{i=1}^{n} x_i\right] + c_1 [n] = \sum_{i=1}^{n} y_i$$

It is interesting to note that if we had a model function of $f(x) = c_3 x^2 + c_2 x + c_1$ and used the calculus method to determine the parameters (it is easier in this case) we would arrive at the system of equations:

$$c_3 \left[\sum_{i=1}^{n} x_i^4\right] + c_2 \left[\sum_{i=1}^{n} x_i^3\right] + c_1 \left[\sum_{i=1}^{n} x_i^2\right] = \sum_{i=1}^{n} x_i^2 y_i$$
$$c_3 \left[\sum_{i=1}^{n} x_i^3\right] + c_2 \left[\sum_{i=1}^{n} x_i^2\right] + c_1 \left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} x_i y_i$$
$$c_3 \left[\sum_{i=1}^{n} x_i^2\right] + c_2 \left[\sum_{i=1}^{n} x_i\right] + c_1 [n] = \sum_{i=1}^{n} y_i$$

Notice the beautiful symmetry in these equations! I bet you could guess the systems of equations if the model function was $f(x) = c_4 x^3 + c_3 x^2 + c_2 x + c_1$!