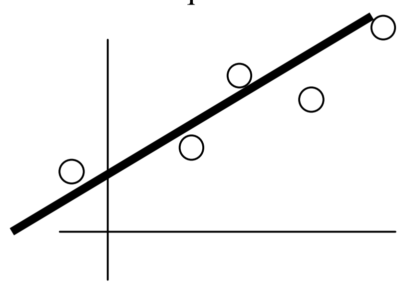
22M:10
Fall 03
J. Simon

Notes on the method of "least squares"
(relates to Section 2.7 in the text)

Suppose you have done some experiment or observations, resulting in a collection of data points (x_1,y_1), (x_2,y_2), ..., (x_n,y_n). You notice that the points seem to lie approximately along a line, and you want to find an equation for a line that fits the data as closely as possible.



A line is specified by 2 parameters, "m" and "b" , for the equation $y=mx+b$. So the mathematical problem is to find the numbers m, b that produce a line that, in some sense, best-fits the data. The text describes one (the most commonly used) way to do this. We can write down a system of two equations in two unknowns (the desired "m" and "b"), and solve the system for m and b.

The text presents this system of two equations in two unknowns in a matrix form. You are welcome to use the book's formula if you want. But I find it easier to understand (and remember) the equations separately (as described below).
***Either way, you end up having to solve a system of two linear equations in two unknowns, so really the work is the same - you are welcome to use either method, depending on which you find easier to understand, remember, and execute.

Here is the system of two equations. It helps me remember them if I think of each of the two equations as saying that the desired "m" and "b" have the property that, with respect to the given data values, (average y) = m(average x) + b. For the first equation, "average" is an unweighted arithmetic average. For the second equation, we are taking weighted averages, with the x-data values as the weights. This kind of averaging is common in statistics, so you probably will encounter it again in our course or subsequent courses.

Equation #1: $(y_1+y_2+...+y_n) / n = m [(x_1+x_2+...+x_n) / n] + b$

Equation #2: $(x_1y_1 + x_2y_2 + ... + x_ny_n) / (x_1+x_2+...+x_n) = m [(x_1x_1+_{x2x2}+...+x_nx_n) / (x_1+x_2+...x_n)] + b$

Remember - this is *one* system of *two* equations; once you set up the equations, you need to solve the system using any of your various methods for solving systems of linear equations. The SOLUTION TO THE SYSTEM is the pair of numbers m,b. The ANSWER to the problem is the "line given by equation y=mx+b" for those particular values of m and b." If you were working some problem in which you found m=7 and b=3 , then the answer to the question, "Which line best fits the given data?" would be "the line y=7x+3".

%%

Additional comment on notation:
In class, I may use a shorter notation to state the two equations.
In Mathematics, Statistics, and various sciences, we often say "add up a bunch of terms" by using the capital greek letter "Sigma", Σ. If we have numbers $x_1, x_2, x_3, ..., x_n$, then I might write "Σx_i , i=1...n" to denote "add up the numbers x_i, where subscript i ranges from 1 to n". If the part "where subscript i ranges from 1 to n" is clear from the context, then I might just write Σx_i. Also, instead of writing x_1x_1, x_2x_2, etc, I might write x_1^2, x_2^2, etc. The equations then are:

Equation #1: $(\Sigma y_i) / n = m [(\Sigma x_i) /n] + b$

Equation #2: $(\Sigma x_i y_i) / (\Sigma x_i) = m [(\Sigma x_i^2) / (\Sigma x_i)] + b$

%%%

Further comment on the form of Equation #2:
I don't like memorizing formulas - I'd much rather have some general principle (that is easier to learn and/or remember) from which particular formulas can be derived. I'm better at deriving than at memorizing. I've written Equation #2 so as to emphasize that it (like Eq 1) is saying some version of "average y = m[average x] + b ". But if fractions make you nervous, you could multiply both sides of Eq 2 by the denominator (Σx_i) to clear the fractions on both sides. That would leave you with :
Equation #2 equivalent : $(\Sigma x_i y_i) = m (\Sigma x_i^2) + b (\Sigma x_i)$

This second form of Equation #2 is a little harder [for me] to remember. But I have to admit that it also is a little better mathematically. In the super-special case where the sum of the x_i values is exactly 0 [e.g. the given points are something like {(-2,1), (-1, 2), (3,4)}] then the "nice" form of Equation #2 is undefined and we have to use the multiplied-out version.

%%%

Here is an example. Suppose we are given the five points
{(1,2), (2,3), (2,4), (3,4), and (5,5)}.
Then

$\Sigma x_i = 1+2+2+3+5 = 13$, $\Sigma y_i = 2+3+4+4+5 = 18$

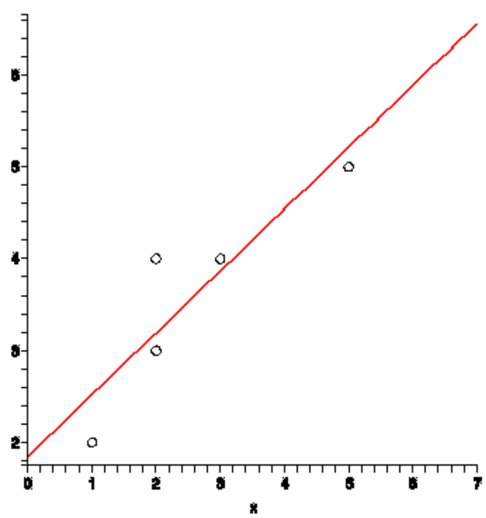
$\Sigma x_i^2 = 1+4+4+9+25 = 43$

$\Sigma x_i y_i = (1)(2) + (2)(3) + (2)(4) + (3)(4) + (5)(5) = 2+6+8+12+25 = 53$,

so the system of equations we want to solve (to find m, b) is

$$\left|\begin{array}{l} (13/5)\ m + b = 18/5 \\ (43/13)\ m + b = 53/13 \end{array}\right.$$

Solving this system, we get
{m = 31/46, b = 85/46}.
So the line y=(31/46)x + (85/46) is the “best fitting” [according to this method] line for the given 5 data points.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[end of handout on Least Squares]