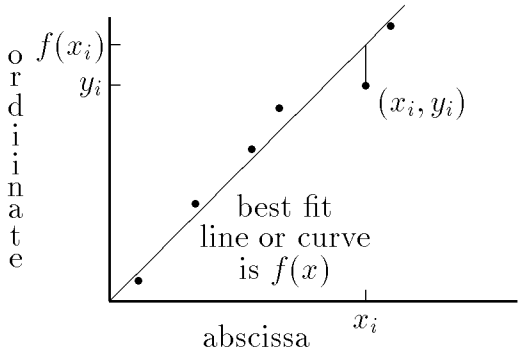
PHSX 221L **Method of Least Squares** **Name:**_____

Object: To obtain a "least squares fit" for different sets of data, using both a linear fit and a parabolic (or second order) fit.

Theory: Often in laboratory work a scientist or engineer makes a series of measurements and the data (x_i, y_i) is displayed by plotting it on a graph. In order to show a correlation or a relationship among the data points it is usually desirable to obtain a line or a curve that will represent the data as closely as possible.

However, measured data contain errors of one type or another which cause the data points to be scattered randomly around some hypothetical line or curve. Thus, in order to find this representative curve we must use some averaging technique. The method of least squares is one such technique. To obtain the curve which best fits the data, we calculate the sum of the squares of the differences between the hypothetical curve and each data point, and then minimize this sum.



If the representative curve is expected to be linear, we write the equation of a line

$$f(x) = Ax + B; \tag{1}$$

if it is thought to be parabolic, then

$$f(x) = Ax^2 + Bx + C. \tag{2}$$

More complicated relationships can be fit to higher-order polynomial curves:

$$f(x) = Ax^n + Bx^{n-1} + Cx^{n-2} + \cdots + Gx + H. \tag{3}$$

In any case, the quantity to be minimized is

$$E = \sum_{i=1}^{N} [f(x_i) - y_i]^2 \tag{4}$$

where y_i is the measured data value (ordinate) corresponding to the abscissa value x_i, and the sum is over the N data points. E is called the least squares error.

The quantity E is a function of the unknown coefficients A, B, C, etc., and therefore to minimize E we take a partial derivative with respect to each of these coefficients and equate the result to zero.

$$\frac{\partial E}{\partial A} = 0, \qquad \frac{\partial E}{\partial B} = 0, \qquad \frac{\partial E}{\partial C} = 0, \ \cdots .$$

This procedure results in a set of simultaneous equations for the unknown coefficients $A, B, C, \ldots$ which can then be solved by standard techniques. Remember, in this procedure, $A, B, C, \ldots$ are the unknowns, not x. Thus, the polynomial function $f(x)$ that best fits the data is found.

If the above procedure is carried out for the straight line case, the two simultaneous equations for A and B in $f(x) = Ax + B$ are:

$$A\sum x_i^2 + B\sum x_i = \sum y_i x_i \tag{5}$$

$$A\sum x_i + B\sum 1 = \sum y_i. \tag{6}$$

Or, in the case of a second order curve:

$$A\sum x_i^4 + B\sum x_i^3 + C\sum x_i^2 = \sum y_i x_i^2 \tag{7}$$

$$A\sum x_i^3 + B\sum x_i^2 + C\sum x_i = \sum y_i x_i \tag{8}$$

$$A\sum x_i^2 + B\sum x_i + C\sum 1 = \sum y_i \tag{9}$$

where each sum is still from $i = 1$ to $i = N$.

Note that while these summations can be performed by hand, they are easier with a modern scientific calculator, and easier still with a spreadsheet program on a computer.

The solution of the system of linear equations can be accomplished by a variety of techniques. You'll recall from algebra class that such systems can be solved by substitution, elimination, row-echelon form with back substitution, Gaussian elimination, Cramer's rule, and matrices. Your calculator probably has a solver for linear systems, and you should be able to get your spreadsheet to invert the appropriate matrix as well. Use these tools for the sums and for the solution to the linear systems in this lab.

Furthermore, your calculator, no doubt, has a built-in linear regression function, and it is performing this least squares calculation; your spreadsheet program probably also has the ability to do linear (and maybe higher order) regressions (again using this method of least squares). However, by learning the details of this method in this lab you are equipped to extend the method to other situations and kinds of functional fits. Once you have learned how this method works in detail you are welcome to use the built-in regression functions of your calculator or spreadsheet on future labs.

Procedure:

1. Use the data from your previous experiment, in which the circumference of a set of disks versus the radius of each disk was measured, to obtain a least squares straight line fit for the plot of circumference versus radius graph that was obtained in that experiment. Make a graph with the original data points and the best fit line. Compare the value you obtain for A with 2π.

2. Suppose you have some compelling theoretical reason to believe that the straight line in procedure 1 *must* go through the origin (*i.e.* $B = 0$). Go through the mathematical procedure described above to arrive at an equation for A if the best fit line is of the form $f(x) = Ax$. Find A (from the same data set) and compare with the A from procedure 1. (For this to work well, you need a data set without a lot of systematic error. A data set with systematic error could give you a better slope if you *don't* force the line through the origin, even if you have reason to believe it should go through the origin.)

3. Do a second order least squares procedure on your mass versus radius data from the same previous experiment to find the equation of the best fit parabola. From your feeling for the physics of the situation, what would you predict B and C to be (before you calculate them)? Compare A with $\rho\pi t_{\text{ave}}$. Make a graph with the original data points and the best fit curve.

4. Do a second order least squares procedure on your displacement versus time data from the previous experiment in which you measured the acceleration of a freely-falling body to find the equation of the best fit parabola. From your feeling for the physics of the situation, what would you predict B and C to be (before you calculate them)? (See equation 2.11 in Serway, replacing x's with y's.) Compare A with $\frac{1}{2}a_g$. Make a graph with the original data points and the best fit curve.

5. Use the following hypothetical set of data to find a least squares straight line and a least squares parabola. Give the values of A, B, C, and E.

x_i	4	12	20	28	32	40	48	60	68	76	80
y_i	2	4	5	6	10	11	16	18	24	26	30

6. Plot both curves and the data from procedure 5 on the same graph. (It would be a good idea to use a computer to plot the functions.) Determine which curve fits the data best (*i.e.* which E is smallest?). The strength of the *linear* association between two variables is often indicated by the correlation coefficient, or Pearson product-moment correlation coefficient, r, which can be expressed as

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{N}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{N}\right)\left(\sum y^2 - \frac{(\sum y)^2}{N}\right)}}. \tag{10}$$

Your calculator or spreadsheet probably reports this correlation coefficient and/or its square (also known as the coefficient of determination).

Discussion:

1. How well do your least squares curves appear to fit the data?
2. What would the least squares polynomial of degree 0 (a horizontal line) be for N points? Interpret.
3. What did you learn from this lab exercise?