Physics 281 LAB 0
DATA ANALYSIS WITH LEAST SQUARES ANALYSIS

For many experiments the graphed data appears to represent a straight line. With some care, the "best" straight line ($y = mx + b$) through the data points can be drawn by eye and from this line the slope (m) and the intercept (b) can be determined. The values of "m" and "b" frequently represent physical quantities that are of interest to us. Because there are errors associated with the data points, we need to find some estimate of the errors in "m" and "b" . This can be doe graphically by drawing two more reasonable lines through the data with slopes greater than and less than that of the first line. These new slopes will give a range to the error of the slope "m" and the new intercepts will give a range to the error in "b". This approach to error analysis does not always give good results.

A more reliable approach uses the method of linear least squares, where all the errors are considered to be randomly distributed in a Gaussian distribution, and there are no errors in the independent variables' measurements. For many experiments the data meet the assumptions for the simple linear least squares method.

If the data appear of the form $y = mx + b$, two sets of measurements (x_1 , y_1) and (x_2,y_2) will determine two equations

$$\mathbf{y_1 = Mx_1 + B} \quad \textbf{(1)}$$

and
$$\mathbf{y_2 = Mx_2 + B}$$

which may be solved to determine M and B. To increase the precision of the results (and to be convinced that the relationship should be linear), a series of measurements are completed with the resulting equations.

$$\mathbf{y_1 = Mx_1 + B} \quad \textbf{(2)}$$
$$\mathbf{y_2 = Mx_2 + B}$$
$$\dots\dots\dots\dots$$
$$\mathbf{y_n = Mx_n + B}$$

Pairs of these equations do not usually determine a unique value for M or B. If we assume there is a unique value for M and B , then the set of equations are no longer equalities and differ from zero by

$$\mathbf{d_1 = y_1 - (mx_1 + b)} \quad \textbf{(3)}$$
$$\mathbf{d_2 = y_2 - (mx_2 + b)}$$
$$\dots\dots\dots\dots\dots\dots$$
$$\mathbf{d_n = y_n - (mx_n + b)}$$

where m and b are now unique and d_1 is the amount y_1 differs from the predicted value $mx_1 + b$. It has now been assumed that there are no errors (or negligibly small errors) in x_1 .

It is apparent that m and b are chosen to minimize the sum of these errors. The sum can be just $d_1 + d_2 + d_3 \text{------} = d_n$,the sum of the absolute values, or any other sum. For linear least squares the sum of the squares of the d_i's is used.

If

$$\mathbf{f} = \Sigma \, (d_i)^2 \quad (4)$$

then m and b are values for which df/dm = 0 and df/db = 0 respectively. Working through the algebra of the equations determined by these derivatives of f, the two equations are solved simultaneously for m and b with the results that:

$$\mathbf{m} = [\ \mathbf{n}(\Sigma \mathbf{xy}) - (\Sigma \mathbf{x})(\Sigma \mathbf{y})\ ]/\mathbf{W} \quad (5)$$

$$\mathbf{b} = [\ (\Sigma \mathbf{x}^2)(\Sigma \mathbf{y}) - (\Sigma \mathbf{xy})(\Sigma \mathbf{x})\ ]/\mathbf{W} \quad (6)$$

with

$$\mathbf{W} = [\ \mathbf{n}(\Sigma \mathbf{x}^2) - (\Sigma \mathbf{x})^2\ ] \quad (7)$$

where the sums such as $\Sigma xy = \Sigma(x_i)(y_i)$ are sums involving all (n) of the data points:

$$\Sigma \mathbf{xy} = \Sigma(\mathbf{x_i})(\mathbf{y_i}) = \mathbf{x_1 y_1} + \mathbf{x_2 y_2} + \mathbf{x_3 y_3} \text{-------------} + \mathbf{x_n y_n} \,. \ (8)$$

The equations for the error in m and b are presented below without derivation. The analysis of the fitting errors depends on an analysis of the origin of errors in a Gaussian distribution. The results for the errors in m and b are as follows:

$$(\Delta \mathbf{m})^2 = \mathbf{n}(\sigma^2)/\mathbf{W} \quad (9)$$

$$(\Delta \mathbf{b})^2 = (\sigma^2)(\Sigma \mathbf{x}^2)/\mathbf{W} \quad (10)$$

$$\sigma^2 = (\Sigma(\mathbf{d_i})^2)/\mathbf{n} = (\Sigma[\mathbf{y_i} - (\mathbf{mx} + \mathbf{b})]^2)/\mathbf{n} \quad (11)$$

The sum of the squares of the deviations (σ^2) calculated above and used to calculate the statistical errors in the slope and intercept (and was the basis for the equations for the "best" slope and "best" intercept) does not include any effects due to the statistical variation in the X variable or due to any systematic errors.

The method of least squares for linear relationships is not restricted to data (x,y) that are related to y = mx + b. The data (u,v) as $v = Au^m$ can be changed to "data" (x,y) that behave as y = mx + b by letting x = ln u, y = ln v, and b = ln A, since $v = Au^m$ is the same as

$$\textbf{ln v = m ln u + ln A} \quad \textbf{(12)}$$

The sums used for a least squares analysis of data that obey a $v = Au^m$ law will be:

$$\Sigma x_i - \Sigma \ln u_i \quad \textbf{(13)}$$

$$\Sigma y_i = \Sigma \ln v_i \quad \textbf{(14)}$$

$$\Sigma(x_i)(y_i) = \Sigma(\ln u_i)(\ln v_i) \quad \textbf{(15)}$$

$$\Sigma(x_i)^2 = \Sigma(\ln u_i)^2 \quad \textbf{(16)}$$

Data (u,v) that obey a mathematical form, $v = Ae^{mu}$, can also be changed to "data" (x,y) that behave as $y = mx + b$, by letting $x = u$, $y = \ln v$, and $b = \ln A$, since $v = Ae^{mu}$ is the same as:

$$\textbf{ln v = mu + ln A} \quad \textbf{(17)}$$

The sums used for a least squares analysis of data that obey a $v = Ae^{mu}$ law will be:

$$\Sigma x_i = \Sigma u_i \quad \textbf{(18)}$$

$$\Sigma y_i = \Sigma \ln v_i \quad \textbf{(19)}$$

$$\Sigma(x_i)(y_i) = \Sigma(u_i)(\ln v_i) \quad \textbf{(20)}$$

$$\Sigma(x_i)^2 = \Sigma(u_i)^2 \quad \textbf{(21)}$$

The linear least squares analysis that will be used in this laboratory mathematically assumes that all the statistical error is in the dependent variable (y), with none in the independent variable. (x**). This more than naive, it is always incorrect!** However for this laboratory we will live with it. The method of least squares for dependent and independent variables with error is nicely presented in an article by William H. Jefferys, in volume 86, pages 177-181 of **The Astrophysical Journal.** We will not use this method in this lab, however as you advance in your studies in science or engineering you most likely will have to use this method.

The slope obtained by linear least squares analysis is actually the weighted average of slopes determined from pairs of data points. The most heavily weighted slope is one determined by a pair of data points, one chosen from each end of the line, $(y_n - y_1)/(x_n - x_1)$.

For more information on linear least squares analysis, see the article by A.M. Bancroft in volume 19, pages 615-616 of the December 1981 issue of **The Physics Teacher**

Another good source on linear least squares analysis (as well as statistics) is the textbook **Data Reduction and Error Analysis for the Physical Sciences** by Phillip Bevington (2nd Edition) 1997. McGraw-Hill

A data fit can be performed by hand or by a computer or calculator (linear regression analysis). Both the high end Hewlett Packard and Texas Instruments calculators have capabilities in this area. (It must be noted that Δm and Δb may not be calculated for you automatically by your calculator, however both calculator brands mentioned above do have the necessary sums stored in statistical data registers so that these calculation can be performed by hand) You also can use MacCurveFit, Excel, Data Logger or another computer statistics program. Normally, this type of program should report m, Δm, b, Δb and r^2. It should be noted that correlation coefficient, R^2, is a measure of the linearity of the data, but is not the directly related to the error in the slope and should not be reported as such. It also gives no indication as to whether the fitted data actually represents what is actually happening physically. R is calculated as follows:

$$r = \frac{[(n\Sigma x_i y_i - \Sigma x_i \Sigma y_i]}{[n\Sigma x_i^2 - (\Sigma x_i)^2]^{1/2} [n\Sigma y_i^2 - (\Sigma y_i)^2]^{1/2}} \quad (22)$$

The value of r ranges from 0, when there is no correlation to +1 or -1, when there is complete correlation. R has the same sign as m.

Three sets of data are given below. Discover their mathematical forms and all statistical errors in any measured quantities. You should start by graphing the data on linear, semi-log, and log-log graph paper to determine which mathematical form best represents the data; linear, power law, or exponential. Then use the proper form of the linear least squares analysis. (see instructions following the data table)Check your results by graphing the equation for the (linear least squares analysis) fit of your data on the same graph paper on which the data is plotted.

SET 1		SET 2		SET 3	
1.0	11.0	1.0	10.87	1.0	3.0
5.5	33.5	4.5	24.46	5.5	38.6
10.2	57.0	6.2	49.27	10.0	94.86
22.5	118.5	7.9	269.7	16.0	192.0
30.1	156.5				

Using the methods outlined previously and using the equations from #4 onward find the least squares best-fit line (The equation that best describes the data set) for all three sets of data. Find the slope and y intercept, the error in the slope and the error in the intercept and the regression coefficient. One data set must be done by hand, using the equations given. (**Show all work**!!!) The other two data sets may be analyzed using your calculator or your computer; if you have statistical or data analysis software. If you use a calculator or software you must mention the name of the calculator or software package used.

This document was originally composed by Dr. Randolph Peterson of the Department of Physics and Astronomy at the University of Tennessee at Chattanooga (Now at the University of the South) in 1985. It was modified and adapted by Harold A. Climer, lab instructor in the Department of Physics, Geology, and Astronomy at the University of Tennessee at Chattanooga 1997/1998/1999/2000/2001.