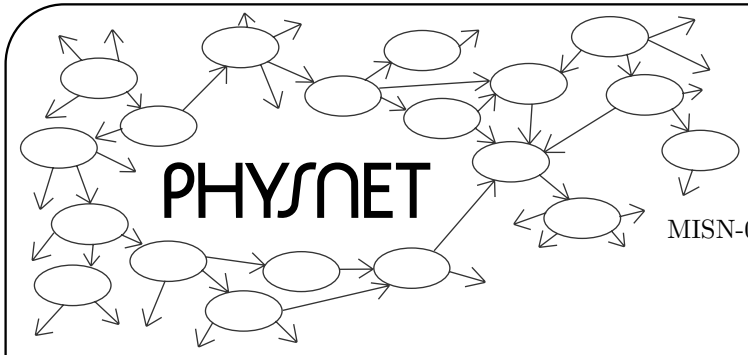
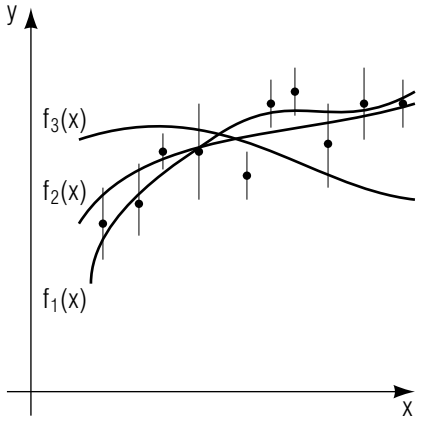


MISN-0-359

LEAST SQUARES FITTING OF EXPERIMENTAL DATA



Project PHYSNET Physics Bldg. Michigan State University East Lansing, MI

LEAST SQUARES FITTING OF EXPERIMENTAL DATA

by

Robert Ehrlich

Title: **Least Squares Fitting of Experimental Data**

Author: R. Ehrlich, Physics Dept., George Mason Univ., Fairfax, VA 22030; (703)323-2303.

Version: 3/22/2002 Evaluation: Stage 0

Length: 2 hr; 24 pages

Input Skills:

1. Vocabulary: cumulative probability, Gaussian distribution, percent error.
2. Enter and run a computer program using advanced programming techniques such as loops and arrays in FORTRAN (MISN-0-347) or in BASIC.
3. Take partial derivatives of simple functions of two variables (MISN-0-201).

Output Skills (Knowledge):

K1. Vocabulary: chi-squared (χ^2), confidence level, correlation term, degrees of freedom, error bars, least squares fit, method of least squares, random error of measurement, residual.

K2. State how least squares fitting is used to determine unknown parameters in a mathematical function used to fit data.

K3. Describe the chi-squared test for the goodness of fit of a mathematical function to a set of experimental data.

Output Skills (Project):

P1. Enter and run a computer program to do a linear least squares fit to a set of experimental or simulated data.

P2. Use the chi-squared value computed for the fit in P1 to determine the goodness of fit and use the computed residuals to reject "bad" data points, then re-run the analysis with the "bad" data points removed.

External Resources (Required):

1. A computer with FORTRAN or BASIC.

LEAST SQUARES FITTING OF EXPERIMENTAL DATA

by

Robert Ehrlich

1. Introduction to Hypothesis Testing

1a. Overview. In this module we discuss and use the "least squares" method for fitting a mathematical function to a set of experimental data. The various values in the data set result from unavoidable experimental error and from controlled variations of one or more independent parameters by the experimentalist. The mathematical fitting function must be a function of the same independent parameters as were varied in the experiment. Such a mathematical fitting function constitutes a hypothesis about how the dependent (measured) quantity varies in nature as the independent variables change. Thus it is a hypothesis as to the theoretical relationship involved. The least squares fitting method is accomplished by varying one or more additional ("free") parameters in the mathematical function so as to minimize the sum of the squares of the deviations of the fitting function from the data at corresponding values of the independent experimentally-varied parameters. The minimum value for the sum of the squared deviations is called "chi-squared" and is written x^2. The values of the free parameters yielding the minimum square deviation sum define the best fit to the experimental data, given the chosen mathematical form for the assumed relationship among the data. A comparison of the value of chi-squared for this fit with a standard chi-squared table enables us to determine the probability that deviations of the data from the theoretical curve could be attributed to the kinds of random fluctuations that are assumed to be present in the particular experimental measurement process. If this "goodness of fit" test yields a low probability of being due solely to the assumed random fluctuations, then it could mean that there are either unexpected errors in the data or that the assumed theoretical relationship is in error. A high probability would only mean that the data are consistent with the theory, not that they and the relationship are necessarily correct.

1b. Consistency Between Data and Theory. In analyzing data from an experiment, we are generally trying to determine whether the data are consistent with a particular theoretical relationship. Suppose the data consist of a set of n values of some measured quantity y: $y_1, y_2, ..., y_n$,

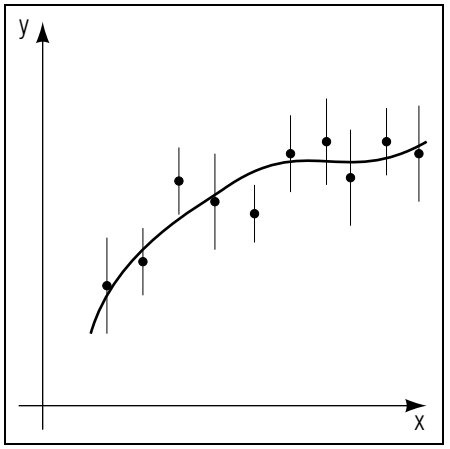


Figure 1. A set of data which is reasonably consistent with function $y = f(x)$.

corresponding to n associated values of some independent variable x: $x_1, x_2, ..., x_n$. In most experiments it is also possible to assign to each data point (x_j, y_j), a "random error" (really an uncertainty) of measurement, σ_j, which depends on the precision of the measuring apparatus. There may also be an uncertainty in the independent x variable, but we shall simplify the analysis by ignoring this possibility. Suppose we wish to test whether x and y satisfy some particular functional relationship $y = f(x)$. A rough simple test of whether the data are consistent with the hypothesized relationship can be made by plotting the function $y = f(x)$ on the same graph as the data $(x_1, y_1), ..., (x_n, y_n)$. The data may be considered roughly consistent with the function $y = f(x)$ if they lie near the curve and are scattered on either side of it in a random looking manner (see Fig. 1). It is highly unlikely that the center points of all of the data will lie exactly on the curve, due to random measurement error: the data and the functional curve shown in Fig. 1 would be considered quite consistent with each other. Note that the vertical "error bar" on each data point extends above and below the point a distance equal to the random measurement "error" σ_j. A data point could be considered "near" the curve if its distance to the curve in the y direction is no greater than the assigned measurement error σ_j, in which case the error bar intersects the curve.This vertical distance from the data point to the curve is the "residual" for the point (x_j, y_j), and is given by

$$r_j = y_j - f(x_j). \tag{1}$$

Note that the residual r_j is positive or negative, depending on whether the data point is above or below the curve.

1c. Consistency Measured In Terms of Probability. The quantity $t_j = r_j/\sigma_j$, is a measure of the discrepancy between the j^th data point and the theoretical curve. We might consider a set of data consistent with a curve even if the magnitudes of some of the residuals r_j are greater than the corresponding measurement errors σ_j, which is the case for two of the data points in Fig. 1. This is because the measurement error σ_j is usually defined such that there is some specific probability (less than 100%) of any particular measurement falling within a range of $\pm\sigma_j$ of the true value. In fact, one way to experimentally determine the measurement error σ_j is to make many repeated measurements of y_j for the same value of x_j, and then find a value for σ_j such that a specific percentage (often chosen to be 68%) of all measurements fall within $\pm\sigma_j$ of the average value y_j. If the theoretical curve $y = f(x)$ is correct, as the number of repeated measurements at the same value of x_j becomes infinite, we expect $y_j = f(x_j)$, i.e., we expect the average of all measurements to lie right on the curve. Furthermore, for any single measurement y_j, there is a 68% chance that the residual r_j is less than σ_j. Thus, there is a 68% chance that $|r_j/^s_j| < 1$ (error bar intersects curve), and a 32% chance that $|r_j/\sigma_j| > 1$ (error bar does not intersect curve).

1d. Chi-Squared and Probability. The quantity chi-squared (x^2) is a measure of the overall discrepancy between all the data points and the theoretical curve, where x^2 is defined as

$$\chi^2 = \sum_{j=1}^{n} t_j^2 = \sum_{j=1}^{n} \left(\frac{r_j}{\sigma_j}\right). \qquad (2)$$

Thus the value of x^2 is a measure of the goodness of fit of a theoretical curve $y = f(x)$ to a set of data points. In the best conceivable fit, the case where all data points would be lying exactly on the curve, we would find $x^2 = 0$. For a reasonably good fit (all the residuals r_j comparable to the measurement errors s_j), we would find $x^2 = n$, where n is the number of data points. If each of the individual measurements is assumed to have a Gaussian probability distribution about some unknown true value it is possible to calculate the probability that x^2 has a particular value. This is usually expressed in terms of a cumulative probability, $P(x^2, n)$ or "confidence level." The number $P(x^2, n)$ is the probability of finding a value for x^2 at least as great as the value actually computed, assuming that the hypothesized function $y = f(x)$ is correct and that the deviations for each data point are only the result of random measurement error. A chi-squared table is given in Table 1.

Table 1. Chi-squared values for sample size and confidence level.

Sample Size (n)	Confidence Level (P)				
	.90	.80	.50	.30	.01
1	.0158	.0642	.455	1.074	6.635
2	.211	.446	1.386	2.408	9.210
3	.584	1.005	2.366	3.665	11.345
4	1.064	1.649	3.357	4.878	13.277
5	1.610	2.343	4.351	6.064	15.086
6	2.204	3.070	5.348	7.231	16.812
7	2.833	3.822	6.346	8.383	18.475
8	3.490	4.594	7.344	9.524	20.090
9	4.168	5.380	8.343	10.656	21.666
10	4.865	6.179	9.342	11.781	23.209

The first column in the table lists n values from 1 to 10. The table entries in the row for each n value are the values of x^2 corresponding to the confidence level listed as column headings. As can be seen for a given value of n, a value of x^2 which greatly exceeds n goes along with a low confidence level. A low confidence level means that the fit is a poor one, and that it is unlikely that the hypothesized relation $y = f(x)$ is consistent with the data.

1e. An Example of Hypothesis Testing. To illustrate these points, suppose we do not know which of three hypothesized functions $f_1(x)$, $f_2(x)$, or $f_3(x)$, describes the relation between the variables y and x (see Fig. 2). For the three functions we use Eqs.(1) and (2) to compute three values of $x^2(x_1^2, x_2^2, x_3^2)$, which indicate the goodness of fit for the three functions. We can use these three values in a chi-squared test in an attempt to determine which of the three functions is correct. Let us assume that the three computed values are $x_1^2 = 6.2$, $x_2^2 = 10.8$, $x_3^2 = 23.2$. We then use the chi-squared table (with $n = 10$) to find the three probabilities corresponding to these values of x^2: $P_1 = .80$, $P_2 = .40$, $P_3 = .01$. An interpolation between table entries is necessary to get P_2. The result $P_3 = .01$ means that only 1% of the time would random errors result in so large a value of x^2, assuming that $y = f_3(x)$ is the correct relation between y and x. We normally regard a confidence level this low as grounds for rejecting the hypothesis that $f_3(x)$ is the correct relation between y and x. For the functions $f_1(x)$ and $f_2(x)$, the respective probabilities 80% and 40% mean that $f_1(x)$ is a somewhat better fit than $f_2(x)$. However, we do not reject $f_2(x)$ simply because the probability P_2 is only half P_1.

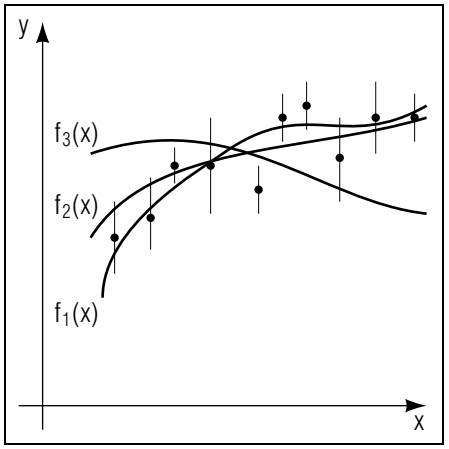


Figure 2. A set of data points and three hypothethized functions $f_1(x)$, $f_2(x)$, and $f_3(x)$. The chi-squared test can be used to reject $f_3(x)$, but it cannot differentiate between $f_1(x)$ and $f_2(x)$.

A 40% confidence level is not so low that the hypothesis can be rejected. Furthermore, we cannot even claim that the probability of $f_1(x)$ being the correct function is twice as great as the probability of $f_2(x)$ being the correct function. The only conclusion we can draw from the chi-squared test is that we cannot decide which of the two functions $f_1(x)$ or $f_2(x)$ is the correct one, since both give reasonably good fits to the data. From this example, we see that the chi-squared test is useful in rejecting hypotheses, but it cannot be used to prove the correctness of a particular hypothesis (unless every other possible hypothesis is rejected).

2. χ^2 Test and Least Squares Fitting

Apart from finding the goodness of fit for comparing specific functions [such as $f_1(x)$, $f_2(x)$ and $f_3(x)$], the chi-squared test can find the goodness of fit for a function f(x) which is not completely specified, but which has a number of "adjustable" parameters $a_1, a_2, ..., a_m$, whose values are not known a priori. In fact, an important application of the chi-squared test is the determination of the values of the parameters which give the best fit to the data. In other words, we wish to determine the values of the parameters $a_1, a_2, ..., a_m$, which give x^2 , the minimum value of the sum of the squares of the residuals. This is known as the method of least squares since it involves minimizing the sum of the squares of deviations (Eq.2) with respect to the adjustable parameters $a_1, a_2, ..., a_m$. The function $f(x)$ which has the minimum square residual sum is called the "least squares fit," and, according to the chi-squared criterion, it is the best fit to the

data. At the same time that we find the least squares fit, we can also determine if the fit is, in fact, a good one. This means finding whether the value of x^2 corresponds to a reasonably high confidence level. In using the chi-squared table to find the confidence level for a fit to n data points which involves m adjustable parameters, we must use the difference $n-m$ (rather than n itself). The difference $n - m$ is known as the number of "degrees of freedom."

3. Linear, One Adjustable Parameter

3a. Formal Solution. The simplest type of least squares fit involves a single adjustable parameter a, which multiplies a function $f(x)$ that is otherwise completely specified: $y = af(x)$. In this case, we may write for S:

$$S = \sum_j \left(\frac{r_j}{\sigma_j}\right)^2 = \sum_j \frac{(y_j - \alpha\, f(x_j))^2}{\sigma_j^2} \tag{3}$$

To find the value of the parameter α which minimizes S, we set the partial derivative $\partial S/\partial\alpha = 0$:

$$\frac{\partial S}{\partial \alpha} = \sum_j \frac{-2f(x_j)(y_j - \alpha f(x_j)}{\sigma_j^2} = 0.$$

Solving for α, we obtain the expression

$$\alpha^* = \frac{\sum_j \dfrac{f(x_j)y_j}{\sigma_j^2}}{\sum_j \dfrac{(f(x_j))^2}{\sigma_j^2}}, \tag{4}$$

where the asterisk indicates that this is the value of α which gives the minimum value of S, which is χ^2. We can obtain an estimate of the uncertainty in α by determining the change in α from the least squares fit value α^* which increases S by some specified amount greater than the minimum value, χ^2. The uncertainty $\Delta\alpha$ is usually defined so that the value of S for both $\alpha = \alpha^* + \Delta\alpha$ and $\alpha = \alpha^* - \Delta\alpha$ is equal to $\chi^2 + 1$, where $\chi^2 = S_{min}$ for $\alpha = \alpha^*$ (see Fig. 3). It can be shown that, for a one-parameter fit, the uncertainty $\Delta\alpha$ is given by

$$\Delta\alpha = \left[\sum_j \frac{(f(x_j))^2}{\sigma_j^2}\right]^{-1/2}. \tag{5}$$

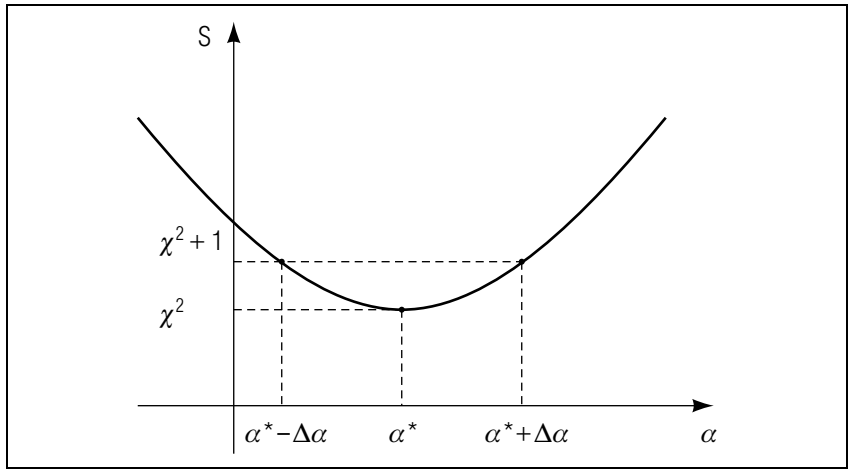


Figure 3. A plot of S against α near the minimum value, chi-square.

A special case of a one-parameter fit that occurs quite frequently is $f(x) = 1$, which gives $y = \alpha$ for the theoretical curve. In this case, the least squares fit value of α^* given by Eq. (4) is the y-intercept of the horizontal straight line which gives the best fit to the data points $(x_1, y_1), ..., (x_n, y_n)$. In effect, α^* is the weighted average of $y_1, ..., y_n$, with weights inversely proportional to the square of the measurement errors $\sigma_1, ..., \sigma_n$.

3b. Example: Fitting ω Mass Data. As an example of the special case $f(x) = 1$, in Table 2 we list nine independent measurements $(y_1, ..., y_9)$ of the mass of the omega meson (a subatomic particle), together with estimated measurement errors $(\sigma_1, ..., \sigma_9)$ from nine different experiments.

Table 2. Values of ω mass and year reported.

Mass (MeV/c^2)	Error (MeV/c^2)	Year Reported
779.4	1.4	1962
784.0	1.0	1963
781.0	2.0	1964
785.6	1.2	1965
786.0	1.0	1966
779.5	1.5	1967
784.8	1.1	1968
784.8	1.1	1969
784.1	1.2	1970

For the nine x-values $x_1, ..., x_9$, we use the year each measurement was reported. Since the mass of a subatomic particle should not depend on the year the measurement was made, the curve we wish to fit to the nine data points, shown plotted in Fig. 4 is a horizontal line $y = \alpha$. To find the value of α which minimizes S, we use Eq.(4) with $f(x) = 1$:

$$\alpha^* = \frac{\sum_j \frac{y_j}{\sigma_j^2}}{\sum_j \frac{1}{\sigma_j^2}}.$$

Upon substitution of the nine masses $(y_1, ..., y_9)$ and measurement errors $(\sigma_1, ..., \sigma_9)$, we find $\alpha^* = 784.12$. Similarly, to find the estimated uncertainty $\Delta\alpha$, we use Eq.(5) with $f(x) = 1$:

$$\Delta\alpha = \left(\sum_j \frac{1}{\sigma_j^2} \right)^{-1/2},$$

from which we find $\Delta\alpha = 0.39$. Thus the value $\alpha = 784.12 + 0.39$ represents the weighted average of the omega meson's mass found from a one-parameter least squares fit to the nine experimental values. The solid horizontal line in Fig. 4 indicates the value of α^* and the dotted horizontal lines indicate the band corresponding to the range $\alpha^* + \Delta\alpha$. In addition to finding the value of one or more parameters from a least squares fit, we can use the chi-squared test to determine if this best fit is a good one. By substitution of $\alpha = 784.12$, $f(x) = 1$, and the experimental values for $y_1, ..., y_9$; $\sigma_1, ..., \sigma_9$, into Eq. (3), we obtain: $\chi^2 = 32.3$. According to the chi-squared table in the Appendix the probability of finding a chi-squared

this high for eight degrees of freedom is less than .01. In other words, the best fit is not a good one in this case. This actually should be obvious from a visual inspection of the data points in Fig. 4, since there is no horizontal line which would pass reasonably close to all the data points.

3c. Whether the "Bad" Data Should be Dropped. If all the data points and measurement errors are correct, the chi-squared test gives grounds for rejecting the hypothesis that the mass of the omega meson is a constant, independent of the year it is measured! A more plausible interpretation of the poor fit is that one or more of the reported measurements is in error. We notice, for example, that if the measurements made in 1962, 1964, and 1967 are dropped, the remaining six measurements are quite consistent with a horizontal line. However, arbitrarily dropping these three data points in order to improve the fit would probably be unwise, because it seems likely that some of the experimenters made systematic (i.e., nonrandom) errors, and statistical tests cannot be used to prove which experiments are in error. In practice, of course, if one experiment out of a large number gives a significantly different result from the others, we might be tempted to discard that result. However, this should not be done unless we can be reasonably certain that the anomalous result cannot be ascribed to better apparatus or other favorable conditions which were absent in all the other experiments, causing them to be in error. Dropping "bad" individual data points within a given experiment is a less dangerous procedure, but even here caution must be exercised lest the "bad" data points reveal systematic errors.

4. Two Adjustable Parameters

4a. Formal Solution. A two-parameter least squares fit to a set of data can be easily made if the relation between y and x is linear in the parameters α_1 and α_2, that is if

$$y = \alpha_1 f_1(x) + \alpha_2 f_2(x),$$

where $f_1(x)$ and $f_2(x)$ are two specified functions. We shall consider the important special case $f_1(x) = 1$ and $f_2(x) = x$, where we wish to make a least squares fit to the straight line $y = \alpha_1 + \alpha_2 x$. In order to minimize the square residual sum with respect to the two parameters a_1 and α_2, we require that the two partial derivatives $\partial S/\partial\alpha_1$ and $\partial S/\partial\alpha_2$ vanish.

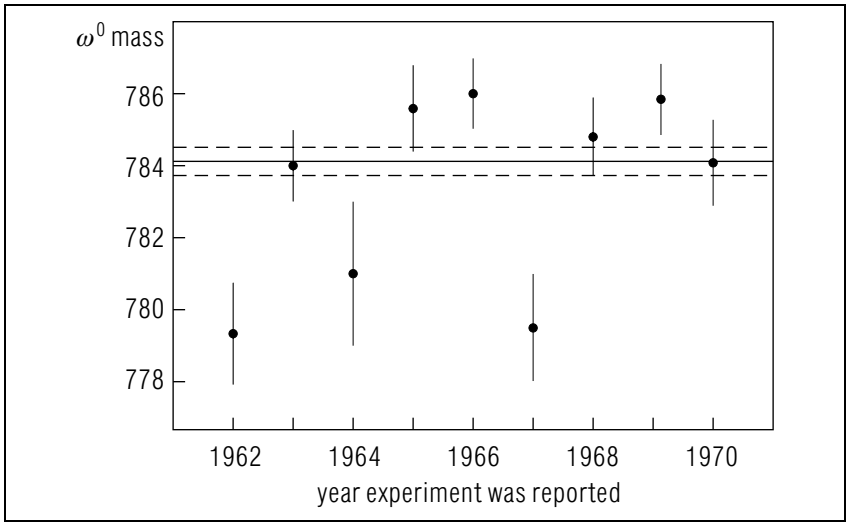


Figure 4. Plot of ω^0 mass against year reported.

Thus, with S given by

$$S = \sum_j \frac{(y_j - \alpha_1 - \alpha_2 x_j)^2}{\sigma_j^2},$$

we require that

$$\frac{\partial S}{\partial \alpha_1} = \sum_j \frac{-2(y_j - \alpha_1 - \alpha_2 x_j)}{\sigma_j^2} = 0 \tag{6}$$

and

$$\frac{\partial S}{\partial \alpha_2} = \sum_j \frac{-2x_j(y_j - \alpha_1 - \alpha_2 x_j)}{\sigma^2} = 0. \tag{7}$$

Solving Eqs.(6) and (7) simultaneously for the two unknowns, α_1 and α_2, we obtain the values of α_1 and α_2 which minimize S:

$$\begin{aligned} \alpha_1^* &= \frac{1}{\Delta}(s_{xx}s_y - s_x s_{xy}) \\ \alpha_2^* &= \frac{1}{\Delta}(s_1 s_{xy} - s_x s_y), \end{aligned} \tag{8}$$

where we have used the abbreviations

$$s_1 = \sum_j \frac{1}{\sigma_j^2};\ s_x = \sum_j \frac{x_j}{\sigma_j^2};\ s_y = \sum_j \frac{y_j}{\sigma_j^2};$$

$$\Delta = s_1 s_{xx} - s_x^2;\ s_{xx} = \sum_j \frac{x_j^2}{\sigma_j^2};\ s_{xy} = \sum_j \frac{x_j y_j}{\sigma_j^2}.$$

As in the case of the one-parameter least squares fit, each of the parameters α_1 and α_2 has an uncertainty which indicates the amount each parameter must be independently changed from the least squares fit values in order to increase the computed value of S to one greater than the minimum value, χ^2. It can be shown that the two uncertainties $\Delta\alpha_1$ and $\Delta\alpha_2$ are given by

$$\begin{aligned} \Delta\alpha_1 &= \left(\frac{s_{xx}}{\Delta}\right)^{1/2} \\ \Delta\alpha_2 &= \left(\frac{s_1}{\Delta}\right)^{1/2} \end{aligned} \tag{9}$$

We may also define a "correlation term" which indicates how the value of χ^2 changes when the parameters are simultaneously varied. In the case of the two-parameter fit, there is one correlation term: $\Delta\alpha_{12} = -s_x/\Delta$.

4b. Pseudo-Linear Least Square Fits. Least squares fits can be found for functions which are not linear but the nonlinear fitting can be quite complex. Therefore it is often desirable to transform the measured variables to other variables which do obey a linear relation. For example, if we measure the level of radioactivity (R) from a radioactive source as a function of time (t) it is found that R is an exponential function of time given by

$$R = R_0 e^{-t/T}. \tag{10}$$

where R_0 is the level of radioactivity at time $t = 0$, and T is the so-called "lifetime," the time for the level of radioactivity to drop to $e^{-1} \cong 37\%$ of its original value. Taking the logarithm of each side of Eq.(10) yields:

$$\ell n\, R = \ell n\, R_0 - t/T. \tag{11}$$

Now substitute $y = \ell n\, R$ and $x = t$ and see that you have transformed the radioactive decay law [Eq.(10)] into a relationship in which the y and x variables are linearly related. This substitution of variables then permits a least squares straight line fit to be done on a set of data from which the value of the lifetime T can be determined. Another reason for

transforming the measured quantities to obtain a linear relation is that we can then often tell by visual inspection whether a straight line is a good fit to the plotted data points, thereby testing the original relational hypothesis.

4c. Example: Acceleration Data. As an illustration of a least squares straight line fit, in Table 3 we show a set of data obtained from a simple experiment using a cart sliding down an incline with negligible friction (an air track). The angle of inclination from the horizontal is $\theta = 0.005546\, rad$. The data represent the measured times for a cart released from rest to travel various distances along the track.

Table 3. Distance (d) vs. time (t) for an air-track cart.

t(sec):	0.7	1.3	1.9	2.5	3.1	3.7	4.1	4.9	5.6
d(cm):	1	4	9	16	25	36	49	64	81
t(sec):	7.5	7.9	8.5	9.1	6.1	6.7			
d(cm):	144	169	196	225	100	121			

We expect the data to be consistent with the relation:

$$d = \frac{1}{2} g \sin\theta\, t^2, \tag{12}$$

where g is the acceleration due to gravity. We can solve Eq.(12) for t to obtain

$$t = \left(\frac{2}{g \sin\theta}\right)^{1/2} d^{1/2}. \tag{13}$$

We can therefore get a linear relation if we make the transformation $t \to y$, $d^{1/2} \to x$. The transformed data points (t against $d^{1/2}$) are plotted in Fig. 5. The small error bars on each point indicate the measurement errors in t. Based on repeated measurements for each point, the error was estimated to be ± 0.1 seconds for all points. The data points seem to be reasonably consistent with a straight line $y = \alpha_1 + \alpha_2 x$, where $\alpha_1 = 0$. In fact, a fairly good straight line fit to the data can be found be simply drawing a straight line with the data points distributed on either side of it in a random manner.

For a more precise result, we can find the slope α_2, y-intercept α_1, and the associated errors $\Delta\alpha_2$ and $\Delta\alpha_1$, for the least squares fit straight line, using Eqs. (8) and (9). We can determine a value for g, the acceleration due to gravity, from the slope of the line. According to the form of

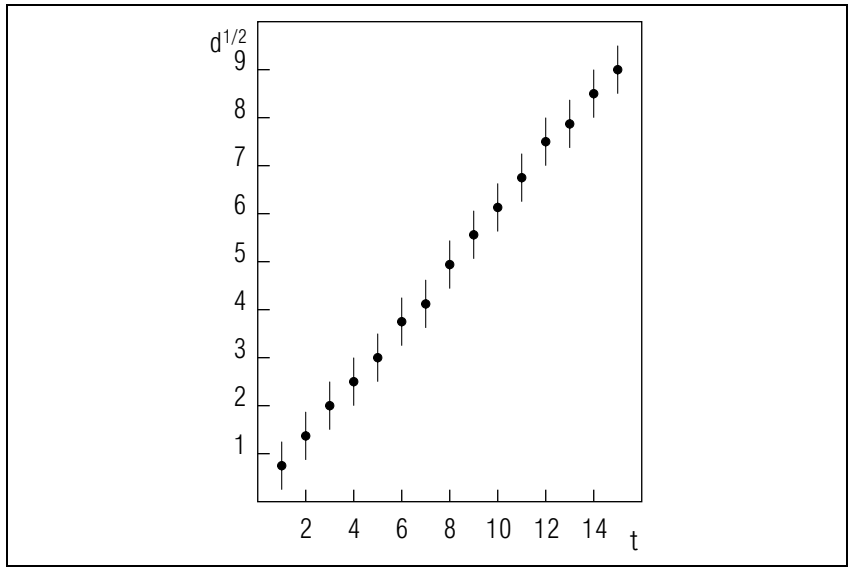


Figure 5. Plot of $d^{1/2}$ against t for data from air-track experiment.

Eq. (13), the slope α_2 is given by

$$\alpha_2 = \left(\frac{2}{g\sin\theta}\right)^{1/2}, \tag{14}$$

which can be solved for g, to obtain

$$g = \frac{2}{\alpha_2^2 \sin\theta}. \tag{15}$$

The theory of first-order error propagation can be used to find the uncertainty in a quantity calculated from a known function of the parameters α_1 and α_2. Since g depends only on α_2, we can find the magnitude of the error in g using

$$\Delta g = \left|\frac{dg}{d\alpha_2}\right| \Delta\alpha_2. \tag{16}$$

From Eqs.(15) and (16), we find that the percentage error in g is twice the percentage error in α_2:

$$\frac{\Delta g}{g} = 2\frac{\Delta\alpha_2}{\alpha_2}. \tag{17}$$

In this example, we could have used our a priori knowledge that the y-intercept should be zero, and found a one-parameter least squares fit to the function $y = \alpha x$. We would use Eqs.(4) and (5) with $f(x) = x$ to obtain

$$\alpha^* = \frac{\sum_j \frac{x_j y_j}{\sigma_j^2}}{\sum_j \frac{x_j^2}{\sigma_j^2}}$$

and

$$\Delta\alpha = \left(\sum_j \frac{x_j^2}{\sigma_j^2}\right)^{-1/2}.$$

One advantage of the one-parameter fit here is that a smaller value for the uncertainty in the slope $\Delta\alpha$ would probably be found due to the more constrained nature of the fit. On the other hand, the two-parameter fit has the advantage that we may possibly establish the existence of a systematic error in the data should we find that α_1^* (the y-intercept) is significantly different from zero.

4d. Fit With Unknown Measurement Errors. We can find a least squares fit even in cases where the measurement errors $\sigma_1, ..., \sigma_n$, are unknown, by assuming that the errors are the same for all points. If we call the common (but unknown) error σ, then Eq.(2) becomes

$$S = \frac{1}{\sigma^2}\sum_{j=1}^{n} r_j^2. \tag{18}$$

Since σ is a constant, we can minimize S with respect to the parameters $\alpha_1, \alpha_2, ..., \alpha_m$, without knowing the value of σ. However, if we wish to determine the uncertainties in the parameters for the best fit: $\Delta\alpha_1, \Delta\alpha_2, ..., \Delta\alpha_m$, a value for σ is needed. A reasonable estimate for σ is the root mean square (RMS) value of the residuals:

$$\left(\frac{\sum r_j^2}{n}\right)^{1/2}, \tag{19}$$

which is a measure of the average "scatter" of the points about the least squares fit curve. Thus it is possible to find values for the parameters $\alpha_1, \alpha_2, ..., \alpha_m$, and estimated uncertainties $\Delta\alpha_1, \Delta\alpha_2, ..., \Delta\alpha_m$, for the least squares fit, even though the measurement errors $\sigma_1, \sigma_2, ..., \sigma_n$ are

unknown. However, it is not possible to use the computed value of χ^2 to tell if the "best" fit is any good because by assuming the value for σ computed via Eq.(19), we have, in effect, assumed a value for χ^2 equal to n, as can be seen by substitution of Eq. (19) into Eq. (18).

5. Program for Least Squares Fits

5a. Input. Our program[1] associated errors (x_j, y_j, σ_j); $j = 1, 2, ..., N$. The errors (σ_j) are optional. The program does a two-parameter least squares straight line fit to the data and it calculates the best fit values of the parameters and their associated errors: $\alpha_1 \pm \Delta\alpha_1$ and $\alpha_2 \pm \Delta\alpha_2$. It also calculates a value for chi-squared for the best-fit straight line. If zero measurement errors are input the program assumes that the σ_j are unknown. In that case, all the σ_j are assumed to be equal to the computed RMS value of the residuals and then chi-squared will not be a meaningful indicator of the goodness of fit.

Table 4. Computer output: Residuals for some air-track data.

NUMBER OF DATA POINTS = 15.00000			
X	Y	DY	RESIDUALS
1.00000	0.70000	0.10000	.02083
2.00000	1.30000	0.10000	.01690
3.00000	1.90000	0.10000	.01298
4.00000	2.50000	0.10000	.00905
5.00000	3.10000	0.10000	.00512
6.00000	3.70000	0.10000	.00119
7.00000	4.10000	0.10000	−.20274
8.00000	4.90000	0.10000	−.00667
9.00000	5.60000	0.10000	.08940
10.00000	6.10000	0.10000	−.01452
11.00000	6.70000	0.10000	−.01845
12.00000	7.50000	0.10000	.17762
13.00000	7.90000	0.10000	−.02631
14.00000	8.50000	0.10000	−.03024
15.00000	9.10000	0.10000	−.03417

5b. Sample Output. The output shown in Table 4 was obtained using the 15 data points listed in Sect.4c. After typing back the input

[1]See this module's Computer Program Supplement.

variables and the calculated residuals, the program outputs the calculated parameters $(\alpha_1, \alpha_2, \chi^2)$ and gives a plot of the best fit straight line, as shown in Fig. 6. The data points are indicated on the plot by the letter "O." Examining the output we notice that the residuals are all reasonably small compared to the measurement errors; in only one case does a residual exceed twice the error. There may, however, be some systematic variation in the residuals, which would indicate the presence of a small systematic error. The plotted least squares fit straight line appears to be quite consistent with the data points (shown as circles). In order to use the table in the Appendix to determine the confidence level for the fit, we need to specify the number of degrees of freedom (13). The value of χ^2 for the least squares fit is 8.5 in the output. According to the χ^2 table, the confidence level for a χ^2 of 8.5 with 13 degrees of freedom is about 0.80, or 80 percent, which means the fit is very good. The values of the parameters α_1, α_2, and their errors, for the least squares fit straight line are given in the output as:

$$\alpha_1 = 0.07523 \pm 0.05433,$$
$$\alpha_2 = 0.60392 \pm 0.00597.$$

The value of the y-intercept (α_1) is therefore almost consistent with zero. The fact that the value is a bit further from zero than the uncertainty $\Delta\alpha_1$ is probably not significant. From the value for the slope α_2, we can compute values for g and Δg using Eqs.(15) and (17):

$$g = 986 \pm 20\,\text{cm/sec}^2,$$

which is consistent with the accepted value, $980\,\text{cm/sec}^2$.

6. A Project

6a. A Linear Least Squares Fit. Run the program using a set of experimental data (see below). The measured quantities in the experiment should either satisfy a linear relationship or should be so related that they can easily be transformed into a pair of linearly related variables. If you have no such data readily available, you may use a simulated set of measurements which might have been obtained from a radioactive decay measurement. Note that the parameters of the fit will be $\alpha_1 = \ell n\, R_0$ and $\alpha_2 = -1/T$ in this case.

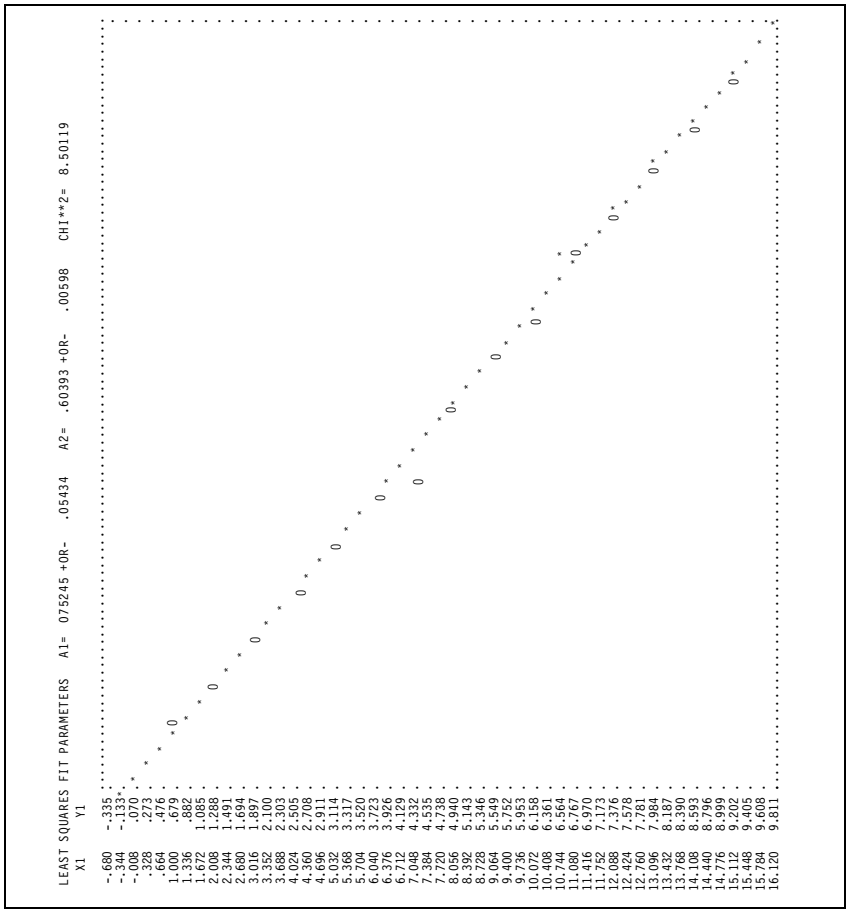


Figure 6. Sample output from the least squares fitting program.

Table 4. Simulated Data from a Radioactive Decay.

t (sec)	$\ell n\, R$	$\Delta(\ell n\, R)$
1.0	10.2	0.2
2.0	9.6	0.2
3.0	8.2	0.2
4.0	5.0	0.2
5.0	6.2	0.2
6.0	4.8	0.2
7.0	4.2	0.2
8.0	2.8	0.2
9.0	2.2	0.2
10.0	3.0	0.2
11.0	−0.6	0.2
12.0	−1.2	0.2

Run the program using this data and use the output to find the lifetime (T) of the radioactive source. Compare your "experimentally" determined lifetime with the "actual" value of $T = 1.0$ seconds.

6b. Goodness-of-Fit Test. Use the computed value of chi-squared for the fit to find a value for the confidence level using the χ^2 table. Based on the computed residuals, are there any very "bad" data points, "bad" in the sense of having residuals large compared to the measurement uncertainty? Redo the fit with these "bad" data points removed. How does the computed lifetime T now agree with the expected value? How good is the fit with the "bad" data points removed, based on the chi-squared table?

Acknowledgments

Preparation of this module was supported in part by the National Science Foundation, Division of Science Education Development and Research, through Grant #SED 74-20088 to Michigan State University.

A. Fortran, Basic, C++ Programs

All programs are at

`http://www.physnet.org/home/modules/support_programs`

which can be navigated to from the home page at

`http://www.physnet.org`

by following the links: → `modules` → `support programs`, where the programs are:

`m359p1f.for`, Fortran;
`m359p1b.bas`, Basic;
`m359p1c.cpp`, C++;
`lib351.h`, needed Library for C++ program;

MODEL EXAM

1-3. See Output Skills K1-K3.

Examinee:
On your computer output sheet(s):

(i) Mark page numbers in the upper right corners of all sheets.

(ii) Label all output, including all axes on all graphs.

On your Exam Answer Sheet(s), for each of the following parts of items (below this box), show:

(i) a reference to your annotated output; and

(ii) a blank area for grader comments.

When finished, staple together your sheets as usual, but include the original of your annotated output sheets just behind the Exam Answer Sheet.

4. Submit your hand-annotated output showing:
   a. the data values;
   b. the residuals and chi-squared for the fit;
   c. a graph of the data values and the fitted curve;
   d. the deduced lifetime;
   e. the standard deviation for the deduced lifetime.
5. Submit your determination of:
   a. the confidence level for the fit;
   b. whether there are any "bad" data;
   c. a refit omitting data hypothesized to be "bad";
   d. a comparison of the expected lifetime with the fitted values, with and without the "bad" data;
   e. a comparison of the confidence levels of the fits, with and without the "bad" data.

INSTRUCTIONS TO GRADER

If the student has submitted copies rather than originals of the computer output, state that on the exam answer sheet and **immediately stop grading the exam and give it a grade of zero.**