CHAPTER 8

Least-Squares Fitting

In our discussion of the statistical analysis of data, we have so far focused exclusively on the repeated measurement of one single quantity, not because the analysis of many measurements of one quantity is the most interesting problem in statistics, but because this simple problem must be well-understood before we can discuss more general ones. We are at last ready to discuss our first, and very important, more general problem.

8.1. *Data that Should Fit a Straight Line*

One of the most common and interesting types of experiment involves the measurement of several values of two different physical variables, in order to investigate the mathematical relationship between the two variables. For instance, one might drop a stone from various different heights $h_1, \ldots, h_N$ and measure the corresponding times of fall $t_1, \ldots, t_N$ to see if the heights and times are connected by the expected relation $h = \frac{1}{2}gt^2$.

Probably the most important experiments of this type are those where the expected relation is *linear*, and this is the case we consider first. For instance, if we believe that a body is falling with constant acceleration g, then its velocity v should be a linear function of the time t,

$$v = v_0 + gt.$$

More generally, we will consider any two physical variables x and y that we suspect are connected by a linear relation of the form

$$y = A + Bx, \tag{8.1}$$

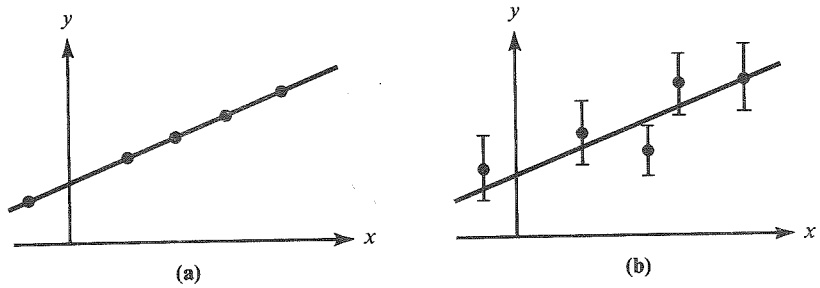

Figure 8.1. (a) If the two variables y and x are linearly related as in Equation (8.1), and if there were no experimental uncertainties, then the measured points (x_i, y_i) would all lie exactly on the line $y = A + Bx$. (b) In practice there always are uncertainties, which can be shown by error bars, and the points (x_i, y_i) can be expected only to lie reasonably close to the line. Here only y is shown as subject to appreciable uncertainties.

where A and B are constants. Unfortunately, many different notations are used for a linear relation; beware of confusing the form (8.1) with the equally popular $y = ax + b$.

If the two variables y and x are linearly related as in (8.1), then a graph of y against x should be a straight line which has slope B and intersects the y axis at $y = A$. If we were to measure N different values $x_1, \ldots, x_N$ and the corresponding values $y_1, \ldots, y_N$, and if our measurements were subject to no uncertainties, then each of the points (x_i, y_i) would lie exactly on the line $y = A + Bx$, as in Figure 8.1(a). In practice, there *are* uncertainties, and the most we can expect is that the distance of each point (x_i, y_i) from the line will be reasonable compared to the uncertainties, as in Figure 8.1(b).

When we make a series of measurements of the kind just described, there are two possible questions we can ask. First, if we take for granted that y and x *are* linearly related, then the interesting problem is to find the straight line $y = A + Bx$ that best fits the measurements; that is, to find the best estimates for the constants A and B based on the data $(x_1, y_1), \ldots, (x_N, y_N)$. This problem can be approached graphically, as discussed briefly in Section 2.6. It can also be treated analytically, by means of the principle of maximum likelihood. This analytical method of finding the best straight line to fit a series of experimental points is called *linear regression*, or the *least-squares fit for a line*, and is the main subject of this chapter.

The second question that can be asked is whether the measured values $(x_1, y_1), \ldots, (x_N, y_N)$ do really bear out our expectation that y is linear in x. We can first find the line that best fits the data, but we must then devise some measure of *how well* this line fits the data. We will take up this second question in Chapter 9.

8.2. *Calculation of the Constants A and B*

Let us now return to the question of finding the best straight line $y = A + Bx$ to fit a set of measured points $(x_1, y_1), \ldots, (x_N, y_N)$. To simplify our discussion, we will suppose that, although our measurements of y suffer some uncertainty, the uncertainty in our measurements of x is negligible. This is often a reasonable assumption, since the uncertainties in one variable often are much larger than those in the other, which we can therefore safely ignore. We will further assume that the uncertainties in y all have the same magnitude. (This is also a reasonable assumption in many experiments, but if the uncertainties are different, then our analysis can be generalized to weight the measurements appropriately; see Problem 8.4.) More specifically, we assume that the measurement of each y_i is governed by the Gauss distribution, with the same width parameter σ_y for all measurements.

If we knew the constants A and B, then, for any given value x_i (which we are assuming has no uncertainty), we could compute the true value of the corresponding y_i,

$$(\text{true value for } y_i) = A + Bx_i. \tag{8.2}$$

The measurement of y_i is governed by a normal distribution centered on this true value, with width parameter σ_y. Therefore, the probability of obtaining the observed value y_i is

$$P_{A,B}(y_i) \propto \frac{1}{\sigma_y} e^{-(y_i - A - Bx_i)^2/2\sigma_y^2}, \tag{8.3}$$

where the subscripts A and B indicate that this probability depends on the (unknown) values of A and B. The probability of obtaining our complete set of measurements $y_1, \ldots, y_N$ is the product

$$\begin{aligned} P_{A,B}(y_1, \ldots, y_N) &= P_{A,B}(y_1) \cdots P_{A,B}(y_N) \\ &\propto \frac{1}{\sigma_y^N} e^{-\chi^2/2}, \end{aligned} \tag{8.4}$$

where the exponent is given by

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}. \tag{8.5}$$

In the now familiar way, the best estimates for the unknown constants A and B, based on the given measurements, are those values of A and B for which the probability $P_{A,B}(y_1, \ldots, y_N)$ is maximum, or for which the sum of squares χ^2 in (8.5) is a minimum (which is why the method is known as least-squares fitting). To find these values, we differentiate χ^2 with respect to A and B and set the derivatives equal to zero:

$$\frac{\partial \chi^2}{\partial A} = (-2/\sigma_y^2) \sum_{i=1}^{N} (y_i - A - Bx_i) = 0 \tag{8.6}$$

and

$$\frac{\partial \chi^2}{\partial B} = (-2/\sigma_y^2) \sum_{i=1}^{N} x_i(y_i - A - Bx_i) = 0. \tag{8.7}$$

These two equations can be rewritten as simultaneous equations for A and B:

$$AN + B\sum x_i = \sum y_i \tag{8.8}$$

and

$$A\sum x_i + B\sum x_i^2 = \sum x_i y_i. \tag{8.9}$$

(From now on we omit the limits $i = 1$ to N from the summation signs $\sum$.) These two equations, known as *normal equations*, are easily solved to give the *least-squares estimates for the constants A and B*

$$A = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{\Delta} \tag{8.10}$$

and

$$B = \frac{N(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\Delta}, \tag{8.11}$$

where we have introduced the convenient abbreviation

$$\Delta = N(\sum x_i^2) - (\sum x_i)^2. \tag{8.12}$$

The results (8.10) and (8.11) give the best estimates for the constants A and B of the straight line $y = A + Bx$, based on the measured points $(x_1, y_1), \ldots, (x_N, y_N)$. The resulting line is called the *least-squares fit* to the data, or the *line of regression* of y on x. It is now natural to ask what are the uncertainties in our estimates for A and B. It turns out that before we can answer this question, we must discuss the uncertainty σ_y in our original measurements of $y_1, \ldots, y_N$, and this we take up next.

8.3. *Uncertainty in the Measurements of y*

In the course of measuring the values $y_1, \ldots, y_N$, we have presumably formed some idea of their uncertainty. Nonetheless, it is important to know how to calculate the uncertainty by analyzing the data themselves. One must remember that the numbers $y_1, \ldots, y_N$ are *not* N measurements of the same quantity. (They might, for instance, be the times for a stone to fall from N different heights.) Thus we certainly do not get an idea of their reliability by examining the spread in their values.

Nevertheless, we can easily estimate the uncertainty σ_y in the numbers $y_1, \ldots, y_N$. The measurement of each y_i is (we are assuming) normally distributed about its true value $A + Bx_i$, with width parameter σ_y. Thus the *deviations* $y_i - A - Bx_i$ are normally distributed, all with the same central value 0 and the same width σ_y. This immediately suggests that a good estimate for σ_y would be given by a sum of squares with the familiar form

$$\sigma_y^2 = \frac{1}{N}\sum (y_i - A - Bx_i)^2. \tag{8.13}$$

In fact, this answer can be confirmed by means of the principle of maximum likelihood. As usual, the best estimate for the parameter in question (σ_y here) is that value for which the probability (8.4) of obtaining the observed values $y_1, \ldots, y_N$ is maximum. As you can easily check by differentiating (8.4) with respect to σ_y and setting the derivative equal to zero, this best estimate is precisely the answer (8.13).

Unfortunately, as you may have suspected, the estimate (8.13) for σ_y^2 is not quite the end of the story. The numbers A and B in (8.13) are the unknown true values of the constants A and B. In practice, these must be replaced by our *best estimates* for A and B, namely, (8.10) and (8.11), and this replacement slightly reduces the value of (8.13). It can be shown that this reduction is compensated for if we replace the factor N in the denominator by $(N-2)$. Thus our final answer for the uncertainty in the measurements $y_1, \ldots, y_N$ is

$$\sigma_y^2 = \frac{1}{N-2}\sum_{i=1}^{N}(y_i - A - Bx_i)^2, \tag{8.14}$$

with A and B given by (8.10) and (8.11). If we already have an independent estimate of our uncertainty in $y_1, \ldots, y_N$, then we would expect this estimate to compare with σ_y as computed from (8.14).

We will not attempt to justify the factor of $(N-2)$ in (8.14), but we can make some comments. First, as long as N is moderately large the difference between N and $(N-2)$ is unimportant anyway. Second, that the factor $(N-2)$ is *reasonable* becomes clear if we consider measuring just two pairs of data (x_1, y_1) and (x_2, y_2). With only two points, we can always find a line that passes *exactly* through both points, and the least-squares fit will give this line. That is, with just two pairs of data, we cannot possibly deduce anything about the reliability of our measurements. Now, since both points lie exactly on the best line, the two terms of the sum in (8.13) and (8.14) are zero. Thus the formula (8.13) (with $N = 2$ in the denominator) would give the absurd answer $\sigma_y = 0$; whereas (8.14), with $N - 2 = 0$ in the denominator, gives $\sigma_y = 0/0$, indicating correctly that σ_y is undetermined after only two measurements.

The presence of the factor $(N-2)$ in (8.14) is reminiscent of the $(N-1)$ that appeared in our estimate of the standard deviation of N measurements of one quantity x, in Equation (5.46). There we made N measurements $x_1, \ldots, x_N$ of the one quantity x. Before we could calculate σ_x, we had to use our data to find the mean $\bar{x}$. In a certain sense, this left only $(N-1)$ independent measured values; so we say that, having computed $\bar{x}$, we have only $(N-1)$ *degrees of freedom* left. Here we made N measurements, but before calculating σ_y we had to compute the *two* quantities A and B. Having done this, we had only $(N-2)$ degrees of freedom left. In general, we define the *number of degrees of freedom* at any stage in a statistical calculation as the number of independent measurements *minus* the number of parameters calculated from these measurements. It is possible to show (though we will not do so here) that it is the number of degrees of freedom, *not* the number of measurements, that should appear in the denominator of formulas like (8.14) and (5.46). This explains why (8.14) contains the factor $(N-2)$, and (5.46) the factor $(N-1)$.

8.4. *Uncertainty in the Constants A and B*

Having found the uncertainty σ_y in the measured numbers $y_1, \ldots, y_N$, we can easily return to our estimates for the constants A and B and calculate their uncertainties. The point is that the estimates (8.10) and (8.11) for A and B are well-defined functions of the measured numbers $y_1, \ldots, y_N$. Therefore the uncertainties in A and B are given by simple error propagation in terms of those in $y_1, \ldots, y_N$. We leave it to the reader to check (Problem 8.8) that

$$\sigma_A^2 = \sigma_y^2 \sum x_i^2/\Delta \tag{8.15}$$

and

$$\sigma_B^2 = N\sigma_y^2/\Delta, \tag{8.16}$$

where Δ is given by (8.12) as usual.

8.5. *An Example*

If the volume of a sample of an ideal gas is kept constant, then its temperature T is a linear function of its pressure P,

$$T = A + BP. \tag{8.17}$$

Here the constant A is the temperature at which the pressure P would drop to zero (if the gas did not condense into a liquid first); it is called the *absolute zero of temperature*, and has the accepted value

$$A = -273.15 \text{ degrees Celsius.} \tag{8.18}$$

The constant B depends on the nature of the gas, its mass, and its volume.[1] By measuring a series of values for T and P, we can find the best estimates for the constants A and B. In particular, the value of A gives the absolute zero of temperature.

One set of five measurements of P and T obtained by a student was as shown in the first three columns of Table 8.1. The student judged that his measurements of P had negligible uncertainty, and those of T were all equally uncertain with an uncertainty of "a few degrees." Assuming that his points should fit a straight line of the form (8.17), he calculated his best estimate for the constant A (the absolute zero) and its uncertainty. What should have been his conclusions?

Table 8.1. Pressure-temperature experiment.

Trial number, i	Pressure, P_i (in mm of mercury)	Temperature, T_i (in °C)	$A + BP_i$
1	65	−20	−22.2
2	75	17	14.9
3	85	42	52.0
4	95	94	89.1
5	105	127	126.2

All we have to do here is use formulas (8.10) and (8.15), with x_i replaced by P_i and y_i by T_i, to calculate all the quantities of interest. This requires us to compute the sums $\sum P_i$, $\sum P_i^2$, $\sum T_i$, $\sum P_i T_i$. Many pocket calculators can evaluate all these sums automatically; but even without such a machine, we can easily handle these calculations if the data are properly organized. From Table 8.1 we can calculate

$$\begin{aligned} \sum P_i &= 425, \\ \sum P_i^2 &= 37{,}125, \\ \sum T_i &= 260, \\ \sum P_i T_i &= 25{,}810, \\ \Delta &= 5{,}000, \end{aligned}$$

where $\Delta = N(\sum P_i^2) - (\sum P_i)^2$. In this kind of calculation, it is important to keep plenty of significant figures, since we have to take differences of these large numbers. Armed with these sums, we can immediately calculate

[1] The difference $T - A$ is called the *absolute temperature*. Thus (8.17) can be rewritten to say that the absolute temperature is proportional to the pressure (at constant volume).

the best estimates for the constants A and B:

$$A = \frac{(\sum P_i^2)(\sum T_i) - (\sum P_i)(\sum P_i T_i)}{\Delta} = -263.35$$

and

$$B = \frac{N(\sum P_i T_i) - (\sum P_i)(\sum T_i)}{\Delta} = 3.71.$$

This already gives the student's best estimate for absolute zero, $A = -263$ degrees Celsius.

Knowing the constants A and B, we can next calculate the numbers $A + BP_i$, the temperatures "expected" on the basis of our best fit to the relation $T = A + BP$. These are shown in the right-hand column of the table, and as we would hope, all agree reasonably well with the observed temperatures. We can now take the difference between the figures in the last two columns and calculate

$$\sigma_T^2 = \frac{1}{N-2} \sum (T_i - A - BP_i)^2 = 44.6$$

and hence the standard deviation,

$$\sigma_T = 6.7.$$

This agrees reasonably with the student's estimate that his temperature measurements were uncertain by "a few degrees."

Finally, we can calculate the uncertainty in A using (8.15):

$$\sigma_A^2 = \sigma_T^2 (\sum P_i^2)/\Delta = 331$$

or

$$\sigma_A = 18.$$

Thus our student's final conclusion, suitably rounded, should be

$$\text{absolute zero, } A = -260 \pm 20 \text{ degrees Celsius,}$$

which agrees satisfactorily with the accepted value, −273 degrees.

As is often true, these results become much clearer if we graph them, as in Figure 8.2. The five data points, with their uncertainties of ±7° in T, are shown on the upper right. The best straight line passes through four of the error bars and close to the fifth.

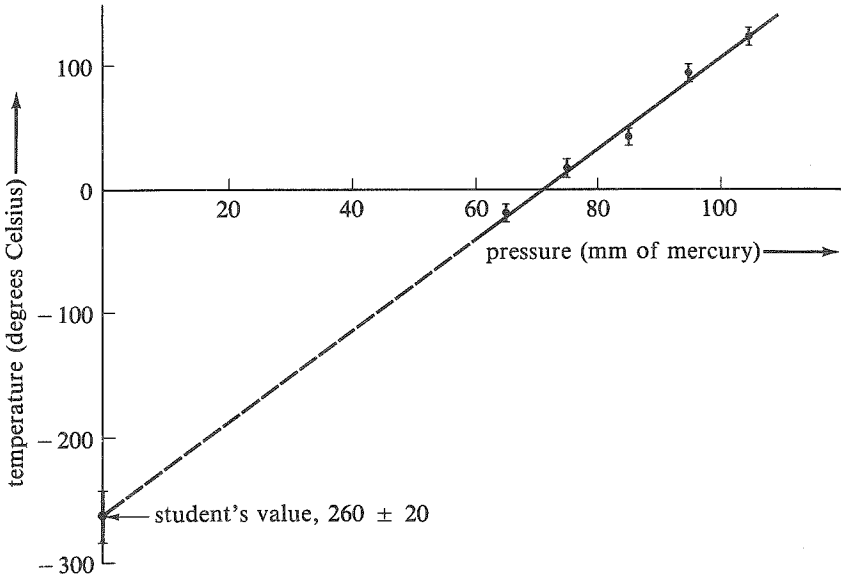

Figure 8.2. Graph of temperature T vs. pressure P for a gas at constant volume. The error bars extend one standard deviation, σ_T, on each side of the five experimental points, and the line is the least-squares best fit. The absolute zero of temperature has been found by extrapolating the line back to its intersection with the T axis.

In order to find a value for absolute zero, the line has been extended beyond all the data points, to its intersection with the T axis. This process of *extrapolation* (extending a curve beyond the data points that determine it) can introduce large uncertainties, as is clear from the picture. A very small change in the line's slope will cause a large change in its intercept on the distant T axis. Thus any uncertainty in the data is greatly magnified if we have to extrapolate any distance. This explains why the uncertainty in the value of absolute zero ($\pm 18°$) is so much larger than that in the original temperature measurements ($\pm 7°$).

8.6. *Least-Squares Fits to Other Curves*

So far in this chapter we have considered the observation of two variables satisfying a linear relation, $y = A + Bx$, and we have discussed the calculation of the constants A and B. This important problem is a special

case of a wide class of curve-fitting problems, many of which can be solved in a similar way. In this last section we mention briefly a few more of these problems.

Fitting a Polynomial

It often happens that one variable, y, is expected to be expressible as a polynomial in a second variable x,

$$y = A + Bx + Cx^2 + \cdots + Hx^n. \tag{8.19}$$

For example, the height y of a falling body is expected to be quadratic in the time t,

$$y = y_0 + v_0 t - \tfrac{1}{2}gt^2,$$

where y_0 and v_0 are the initial height and velocity, and g is the acceleration of gravity. Given a set of observations of the two variables, one can find best estimates for the constants $A, B, \ldots, H$ in (8.19) by an argument which exactly parallels that of Section 8.2, as we now outline.

To simplify matters, we suppose that the polynomial (8.19) is actually a quadratic,

$$y = A + Bx + Cx^2. \tag{8.20}$$

(The interested reader can easily extend the analysis to the general case.) We suppose, as before, that we have a series of measurements (x_i, y_i), $i = 1, \ldots, N$, with the y_i all equally uncertain and the x_i all exact. For each x_i, the corresponding true value of y_i is given by (8.20), with A, B, and C as yet unknown. We assume that the measurements of the y_i are governed by normal distributions, each centered on the appropriate true value and all with the same width σ_y. This lets us compute the probability of obtaining our observed values $y_1, \ldots, y_N$ in the familiar form

$$P(y_1, \ldots, y_N) \propto e^{-\chi^2/2}, \tag{8.21}$$

where now

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i - Cx_i^2)^2}{\sigma_y^2}. \tag{8.22}$$

(This corresponds to Equation (8.5) for the linear case.) The best estimates for A, B, and C are those values for which $P(y_1, \ldots, y_N)$ is largest, or χ^2 is smallest. Differentiating χ^2 with respect to A, B, and C and setting these derivatives equal to zero, we obtain the three equations (as you should check):

$$\begin{aligned} AN + B\sum x_i + C\sum x_i^2 &= \sum y_i, \\ A\sum x_i + B\sum x_i^2 + C\sum x_i^3 &= \sum x_i y_i, \\ A\sum x_i^2 + B\sum x_i^3 + C\sum x_i^4 &= \sum x_i^2 y_i. \end{aligned} \tag{8.23}$$

For any given set of measurements (x_i, y_i), these simultaneous equations for A, B, and C (known as the *normal equations*) can be solved to find the best estimates for A, B, and C. With A, B, C calculated in this way, the equation $y = A + Bx + Cx^2$ is called the least-squares polynomial fit, or the polynomial regression, for the given measurements.

The method of polynomial regression generalizes easily to a polynomial of any degree, although the resulting normal equations become very cumbersome for polynomials of high degree. In principle, a similar method can be applied to *any* function $y = f(x)$ that depends on various unknown parameters $A, B, \ldots$. Unfortunately, the resulting normal equations that determine the best estimates for $A, B, \ldots$ can be difficult or impossible to solve. However, there is one large class of problems which *can* always be solved, namely, those problems where the function $y = f(x)$ depends linearly on the parameters $A, B, \ldots$. These include all polynomials—obviously the polynomial (8.19) is linear in its coefficients $A, B, \ldots$—but they include many other functions. For example, in some problems y is expected to be a sum of trigonometric functions, like

$$y = A \sin x + B \cos x. \tag{8.24}$$

For this function, and in fact for any function that is linear in the parameters $A, B, \ldots$, the normal equations that determine the best estimates for $A, B, \ldots$ are simultaneous linear equations, which can always be solved (see Problems 8.12 and 8.13).

Exponential Functions

One of the most important functions in physics is the exponential function

$$y = Ae^{Bx}, \tag{8.25}$$

where A and B are constants. The intensity I of radiation, after passing a distance x through a shield, falls off exponentially:

$$I = I_0 e^{-\mu x},$$

where I_0 is the original intensity and μ characterizes the absorption by the shield. The charge on a short-circuited capacitor drains away exponentially:

$$Q = Q_0 e^{-\lambda t}$$

where Q_0 is the original charge and $\lambda = 1/(RC)$, R and C being the resistance and capacitance.

If the constants A and B in (8.25) are unknown, then it is natural to seek estimates of them based on measurements of x and y. Unfortunately, direct application of our previous arguments leads to equations for A and B that cannot be conveniently solved. However, it is possible to transform the nonlinear relation (8.25) between y and x into a linear relation, to which we can apply our least-squares fit.

To effect the desired "linearization," we simply take the logarithm of (8.25) to give

$$\ln y = \ln A + Bx. \tag{8.26}$$

We see that, even though y is not linear in x, $\ln y$ *is*. This conversion of the nonlinear (8.25) into the linear (8.26) is useful in many contexts besides that of least-squares fitting. If we wish to check the relation (8.25) graphically, then a direct plot of y against x will produce a curve that is hard to identify visually. On the other hand, a plot of $\ln y$ against x (or of $\log y$ against x) should produce a straight line, which can be easily identified. (Such a plot is especially easy if one uses "semilog" graph paper, on which the graduations on one axis are spaced logarithmically. Such paper lets one plot $\log y$ directly without even calculating it.)

The usefulness of the linear equation (8.26) in least-squares fitting is readily apparent. If we believe that y and x should satisfy $y = Ae^{Bx}$, then the variables $z = \ln y$ and x should satisfy (8.26), or

$$z = \ln A + Bx. \tag{8.27}$$

If we have a series of measurements (x_i, y_i), then for each y_i we can calculate $z_i = \ln y_i$. Then the pairs (x_i, z_i) should lie on the line (8.27). This line can be fitted by the method of least squares to give best estimates for the constants $\ln A$ (from which we can find A) and B.

Example

Many populations (of people, of bacteria, of radioactive nuclei, etc.) tend to vary exponentially in time. If a population N is decreasing exponentially, we write

$$N = N_0 e^{-t/\tau}, \tag{8.28}$$

where τ is called the population's *mean life* (closely related to the *half-life*, $t_{1/2}$; in fact, $t_{1/2} = 0.693\tau$). A biologist suspects that a population of bacteria is decreasing exponentially as in (8.28), and measures the population on three successive days, with the results shown in the first two columns of Table 8.2. Given these data, what is his best estimate for the mean life τ?

Table 8.2. Population of bacteria.

Time t_i (days)	Population N_i	$z_i = \ln N_i$
0	153,000	11.94
1	137,000	11.83
2	128,000	11.76

If N varies as in (8.28), then the variable $z = \ln N$ should be linear in t:

$$z = \ln N = \ln N_0 - \frac{t}{\tau}. \tag{8.29}$$

Our biologist therefore calculates the three numbers $z_i = \ln N_i$ $(i = 0, 1, 2)$ shown in the third column of Table 8.2. Using these numbers, he makes a least-squares fit to the straight line (8.29) and finds as best estimates for the coefficients $\ln N_0$ and $(-1/\tau)$,

$$\ln N_0 = 11.93 \quad \text{and} \quad (-1/\tau) = -0.089 \text{ day}^{-1}.$$

The second of these implies that his best estimate for the mean life is

$$\tau = 11.2 \text{ days}.$$

The method just described is attractively simple (especially with a calculator that performs linear regression automatically) and is frequently used. Nevertheless, the method is not quite logically sound. Our derivation of the least-squares fit to a straight line $y = A + Bx$ was based on the assumption that the measured values $y_1, \ldots, y_N$ were all equally uncertain. Here we are performing our least-squares fit using the variable $z = \ln y$. Now, if the measured values y_i are all equally uncertain, then the values $z_i = \ln y_i$ are *not*. In fact, from simple error propagation we know that

$$\sigma_z = \left|\frac{dz}{dy}\right| \sigma_y = \frac{\sigma_y}{y}. \tag{8.30}$$

Thus if σ_y is the same for all measurements, then σ_z varies (with σ_z larger when y is smaller). Evidently, the variable $z = \ln y$ does not satisfy the requirement of equal uncertainties for all measurements, if y itself does.

The remedy for this difficulty is straightforward. One can modify the least-squares procedure to allow for different uncertainties in the measurements, provided the various uncertainties are known. (This method of *weighted least squares* is outlined in Problem 8.4.) If we know that the measurements of $y_1, \ldots, y_N$ really are equally uncertain, then Equation (8.30) tells us how the uncertainties in $z_1, \ldots, z_N$ vary, and we can therefore apply the method of weighted least squares to the equation $z = \ln A + Bx$.

In practice, one often cannot be sure that the uncertainties in $y_1, \ldots, y_N$ really are constant; so one can perhaps argue that one could just as well assume the uncertainties in $z_1, \ldots, z_N$ to be constant and use the simple, unweighted least squares. Often the variation in the uncertainties is small, and it makes little difference which method is used, as was true in the preceding example. In any event, straightforward application of the ordinary (unweighted) least-squares fit is an unambiguous and simple way to get *reasonable* (if not *best*) estimates for the constants A and B in the equation $y = Ae^{Bx}$; so it is frequently used in this way.

Multiple Regression

Finally, we have so far discussed only observations of *two* variables, x and y, and their relationship. In many real problems there are more than two variables to be considered. For example, in studying the pressure P of a gas, one finds that it depends on the volume V and temperature T, and one must analyze P as a function of V and T. The simplest example of such a problem is when one variable, z, depends linearly on two others, x and y:

$$z = A + Bx + Cy. \tag{8.31}$$

This problem can be analyzed by a very straightforward generalization of our two-variable method. If we have a series of measurements (x_i, y_i, z_i), $i = 1, \ldots, N$ (with the z_i all equally uncertain, and the x_i and y_i exact) then we can use the principle of maximum likelihood exactly as in Section

8.2 to show that the best estimates for the constants A, B, C are determined by normal equations of the form

$$\begin{aligned} AN + B\sum x_i + C\sum y_i &= \sum z_i, \\ A\sum x_i + B\sum x_i^2 + C\sum x_i y_i &= \sum x_i z_i, \\ A\sum y_i + B\sum x_i y_i + C\sum y_i^2 &= \sum y_i z_i. \end{aligned} \tag{8.32}$$

The equations can be solved for A, B, and C to give the best fit for the relation (8.31). This method is called *multiple regression* ("multiple" since there are more than two variables), but we will not discuss it further here.

Problems

Reminder: An asterisk indicates that the problem is discussed, or its answer given, in the Answers section at the back of the book.

***8.1** (Section 8.2). Use the method of least squares to find the line $y = A + Bx$ that best fits the four points (1, 12), (2, 13), (3, 18), (4, 19). Plot the points and line.

8.2 (Section 8.2). To find the spring constant k of a spring, a student loads it with various masses m and measures the corresponding lengths l. Her results are shown in Table 8.3.

Table 8.3.

load m (gm)	200	300	400	500	600	700	800	900
length l (cm)	5.1	5.5	5.9	6.8	7.4	7.5	8.6	9.4

Since the force mg is $k(l - l_0)$, where l_0 is the unstretched length of the spring, these data should fit the line $l = l_0 + (g/k)m$. Make a least-squares fit to the data, and find the best estimates for the unstretched length l_0 and the spring constant k.

***8.3** (Section 8.2). Suppose two variables x and y are known to satisfy the relation $y = Bx$; i.e., they lie on a straight line that is known to pass through the origin. Suppose further that you have N measurements (x_i, y_i), with the uncertainties in x negligible and those in y all equal. Using arguments like those in Section 8.2, show that the least-squares best estimate for B is

$$B = \sum x_i y_i / \sum x_i^2.$$

***8.4** (Section 8.2). Suppose we measure N pairs of values (x_i, y_i) of two variables x and y that are supposed to satisfy a linear relation $y = A + Bx$. Suppose the measurements of the x_i have negligible uncertainty, and those of the y_i have different uncertainties σ_i. (That is, y_1 has uncertainty σ_1, while y_2 has uncertainty σ_2, and so on.) Review the derivation of the least-squares fit in Section 8.2, and then generalize it to cover this situation where the uncertainties in the y_i are not all the same. Show that the best estimates of A and B are

$$A = [(\sum w_i x_i^2)(\sum w_i y_i) - (\sum w_i x_i)(\sum w_i x_i y_i)]/\Delta \tag{8.33}$$

and

$$B = [(\sum w_i)(\sum w_i x_i y_i) - (\sum w_i x_i)(\sum w_i y_i)]/\Delta, \tag{8.34}$$

with weights $w_i = 1/\sigma_i^2$ and

$$\Delta = (\sum w_i)(\sum w_i x_i^2) - (\sum w_i x_i)^2. \tag{8.35}$$

This method of *weighted least squares* can be applied only when the uncertainties σ_i (or at least their relative sizes) are known. Perhaps the commonest situation where this is so is a counting experiment, like the counting of radioactive decays. As discussed in Section 3.1 (and proved in Chapter 11), the uncertainty corresponding to any count v is known to be $\sqrt{v}$.

8.5 (Section 8.2). Suppose y is known to be linear in x, so that $y = A + Bx$, and suppose we have three measurements of (x, y): $(1, 2 \pm .5)$; $(2, 3 \pm .5)$; $(3, 2 \pm 1.5)$, for which the uncertainties in x are negligible. Use the method of weighted least squares, Equations (8.33) to (8.35), to calculate A and B. Compare your results with what you would get if you ignored the variation in the uncertainties, i.e., used the unweighted fit of Equations (8.10) to (8.12). Plot the data and both lines, and try to understand the differences.

***8.6** (Section 8.4). A train, presumed to be traveling at constant speed, is timed as it goes past four different positions, with the results shown in Table 8.4. By making a least-squares fit to the line $d = d_0 + vt$, find the best estimate for the train's speed, v. What is the uncertainty in v?

Table 8.4.

distance (feet)	0	3000	6000	9000
time (seconds)	17.6	40.4	67.7	90.1

8.7 (Section 8.4). A student measures the pressure P of a gas at five different temperatures T, keeping the volume V fixed. His results are shown in Table 8.5.

Table 8.5.

pressure P_i (mm of mercury)	79	82	85	88	90
temperature T_i (°Celsius)	8	17	30	37	52

His data should fit a linear equation of the form $T = A + BP$, where A is the absolute zero of temperature (whose accepted value is $-273°$ Celsius, as discussed in Section 8.5). Find the best fit to the student's data, and hence his best estimate for absolute zero and its uncertainty.

***8.8** (Section 8.4).
(a) Use the principle of maximum likelihood, as outlined in the discussion of Equation (8.13), to show that (8.13) gives the uncertainty σ_y of y in a series of measurements $(x_1, y_1), \ldots, (x_N, y_N)$ that are supposed to fit a straight line.
(b) Use error propagation to show that the uncertainties σ_A and σ_B in the parameters of a straight line $y = A + Bx$ are given by (8.15) and (8.16).

***8.9** (Section 8.4). The least-squares fit to a set of points $(x_1, y_1), \ldots, (x_N, y_N)$ treats the variables x and y unsymmetrically. Specifically, one finds a best fit for the line $y = A + Bx$ by assuming that the numbers $y_1, \ldots, y_N$ are all equally uncertain but that $x_1, \ldots, x_N$ have negligible uncertainty. If the situation were reversed, then one would have to interchange the roles of x and y and fit to a line $x = A' + B'y$. The two lines $y = A + Bx$ and $x = A' + B'y$ would be the same if the N points lie *exactly* on a line, but in general the two lines will be slightly different. Fit the data of Problem 8.1 to a line $x = A' + B'y$ (treating the x_i as equally uncertain and the y_i as certain). Find A' and B' and their uncertainties $\sigma_{A'}$ and $\sigma_{B'}$. What would be the values of A' and B' based on the answers to Problem 8.1? Compare the lines found by the two methods. Is the difference significant?

8.10 (Section 8.6). Consider the problem of fitting a set of measurements (x_i, y_i), $i = 1, \ldots, N$, to the polynomial $y = A + Bx + Cx^2$. Use the principle of maximum likelihood to show that the best estimates for A, B, C based on the data are given by Equations (8.23). Follow the arguments outlined between Equations (8.20) and (8.23).

***8.11** (Section 8.6). One way to measure the acceleration of a freely falling body is to measure its height y_i at a succession of equally spaced times t_i (with a multiflash photograph, for example) and to find the best fit to the expected polynomial

$$y = y_0 + v_0 t - \tfrac{1}{2} g t^2. \tag{8.36}$$

Use the equations (8.23) to find the best estimates for the three coefficients in (8.36), and hence the best estimate for g, based on the five measurements in Table 8.6.

Table 8.6.

time t (tenths of sec)	-2	-1	0	1	2
height y (cm)	131	113	89	51	7

Note that we can name the times however we like. A more natural choice might seem to be $t = 0, 1, \ldots, 4$. However, when you solve the problem you will see that defining the times to be symmetrically spaced about zero causes about half the sums involved to be zero and greatly simplifies the calculations. This trick can be used whenever the values of the independent variable are equally spaced.

8.12 (Section 8.6). Suppose that y is expected to have the form $y = Af(x) + Bg(x)$, where A and B are unknown parameters, and f and g are fixed, known functions (such as $f = x$ and $g = x^2$, or $f = \cos x$ and $g = \sin x$). Use the principle of maximum likelihood to show that the best estimates for A and B, based on data (x_i, y_i), $i = 1, \ldots, N$, must satisfy

$$\begin{aligned} A\textstyle\sum[f(x_i)]^2 + B\sum f(x_i)g(x_i) &= \sum y_i f(x_i), \\ A\textstyle\sum f(x_i)g(x_i) + B\sum[g(x_i)]^2 &= \sum y_i g(x_i). \end{aligned} \tag{8.37}$$

***8.13** (Section 8.6). A weight oscillating on a vertical spring should have height y given by

$$y = A \cos \omega t + B \sin \omega t.$$

A student measures ω to be 10 rad/sec with negligible uncertainty. Using a multiflash photograph, she then finds y for five equally spaced times, as shown in Table 8.7.

Table 8.7.

t (tenths of sec)	-4	-2	0	2	4
y (cm)	3	-16	6	9	-8

Use Equations (8.37) to find best estimates for A and B. Plot the data and your best fit. (If you plot the data first, you will have the opportunity to consider how hard it would be to choose a best fit without the least-squares method.) If the student judges that her measured values of y were uncertain by "a couple of centimeters," would you say that the data are an acceptable fit to the expected curve?

***8.14** (Section 8.6). The rate R at which a sample of radioactive material emits radiation decreases exponentially as the material is depleted:

$$R = R_0 e^{-t/\tau},$$

where τ is the mean life of the sample. A student observed a certain radioactive material for three hours with the results shown in Table 8.8. By making a least-squares fit to the line $\ln R = \ln R_0 - t/\tau$, find the best estimate for the mean life τ.

Table 8.8.

time t (hours)	0	1	2	3
rate R (arbitrary units)	13.8	7.9	6.1	2.9