Chapter 8

Least-Squares Fitting

Our discussion of the statistical analysis of data has 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 more general ones can be discussed. Now we are 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 to investigate the mathematical relationship between the two variables. For instance, an experimenter 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 for which the expected relation is *linear*. 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

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

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 that 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

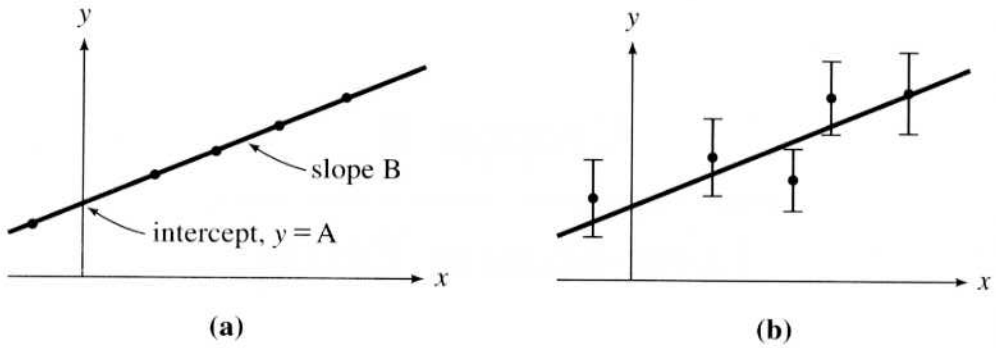

Figure 8.1. **(a)** If the two variables y and x are linearly related as in Equation (8.1), and if there were no experimental uncertainties, then the measured points (x_i, y_i) would all lie exactly on the line $y = A + Bx$. **(b)** In practice, there always are uncertainties, which can be shown by error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

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 with the uncertainties, as in Figure 8.1(b).

When we make a series of measurements of the kind just described, we can ask two questions. 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. To answer this question, we would first find the line that best fits the data, but we must then devise some measure of *how well* this line fits the data. If we already know the uncertainties in our measurements, we can draw a graph, like that in Figure 8.1(b), that shows the best-fit straight line and the experimental data with their error bars. We can then judge visually whether or not the best-fit line passes sufficiently close to all of the error bars. If we do not know the uncertainties reliably, we must judge how well the points fit a straight line by examining the distribution of the points themselves. We take up this 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 appreciable uncertainty, the uncertainty in our measurements of x is negligible. This assumption is often reasonable, because 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 assumption is also reasonable in many experiments, but if the uncertainties are different, then our analysis can be generalized to weight the measurements appropriately; see Problem 8.9.) 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

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

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

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

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, we will assume that 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 $Prob_{A,B}(y_1, \ldots, y_N)$ is maximum, or for which the sum of squares χ^2 in (8.5) is a minimum. (This 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:

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

and

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

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

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

and

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

Here, I have omitted the limits $i = 1$ to N from the summation signs Σ. In the following discussion, I also omit the subscripts i when there is no serious danger of confusion; thus, $\Sigma x_i y_i$ is abbreviated to Σxy and so on.

The two equations (8.8) and (8.9), sometimes called *normal equations,* are easily solved for the least-squares estimates for the constants A and B,

$$A = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{\Delta} \tag{8.10}$$

and

$$B = \frac{N\Sigma xy - \Sigma x \Sigma y}{\Delta}, \tag{8.11}$$

where I have introduced the convenient abbreviation for the denominator,

$$\Delta = N\Sigma x^2 - (\Sigma x)^2. \tag{8.12}$$

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 N 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.

Example: Length versus Mass for a Spring Balance

A student makes a scale to measure masses with a spring. She attaches its top end to a rigid support, hangs a pan from its bottom, and places a meter stick behind the arrangement to read the length of the spring. Before she can use the scale, she must calibrate it; that is, she must find the relationship between the mass in the pan and the length of the spring. To do this calibration, she gets five accurate 2-kg masses, which she adds to the pan one by one, recording the corresponding lengths l_i as shown in the first three columns of Table 8.1. Assuming the spring obeys Hooke's law, she anticipates that l should be a linear function of m,

$$l = A + Bm. \tag{8.13}$$

(Here, the constant A is the unloaded length of the spring, and B is g/k, where k is the usual spring constant.) The calibration equation (8.13) will let her find any unknown mass m from the corresponding length l, once she knows the constants A and B. To find these constants, she uses the method of least squares. What are her answers for A and B? Plot her calibration data and the line given by her best fit (8.13). If she puts an unknown mass m in the pan and observes the spring's length to be $l = 53.2$ cm, what is m?

Table 8.1. Masses m_i (in kg) and lengths l_i (in cm) for a spring balance. The "x" and "y" in quotes indicate which variables play the roles of x and y in this example.

Trial number i	"x" Load, m_i	"y" Length, l_i	m_i^2	$m_i l_i$
1	2	42.0	4	84
2	4	48.4	16	194
3	6	51.3	36	308
4	8	56.3	64	450
5	10	58.6	100	586
$N = 5$	$\Sigma m_i = 30$	$\Sigma l_i = 256.6$	$\Sigma m_i^2 = 220$	$\Sigma m_i l_i = 1{,}622$

As often happens in such problems, the two variables are not called x and y, and one must be careful to identify which is which. Comparing (8.13) with the standard form, $y = A + Bx$, we see that the length l plays the role of the dependent variable y, while the mass m plays the role of the independent variable x. The constants A and B are given by (8.10) through (8.12), with the replacements

$$x_i \leftrightarrow m_i \qquad \text{and} \qquad y_i \leftrightarrow l_i.$$

(This correspondence is indicated by the headings "x" and "y" in Table 8.1.) To find A and B, we need to find the sums Σm_i, Σl_i, Σm_i^2, and $\Sigma m_i l_i$; therefore, the last two columns of Table 8.1 show the quantities m_i^2 and $m_i l_i$, and the corresponding sum is shown at the bottom of each column.

Computing the constants A and B is now straightforward. According to (8.12), the denominator Δ is

$$\begin{aligned} \Delta &= N\Sigma m^2 - (\Sigma m)^2 \\ &= 5 \times 220 - 30^2 = 200. \end{aligned}$$

Next, from (8.10) we find the intercept (the unstretched length)

$$\begin{aligned} A &= \frac{\Sigma m^2 \Sigma l - \Sigma m \Sigma ml}{\Delta} \\ &= \frac{220 \times 256.6 - 30 \times 1622}{200} = 39.0 \text{ cm}. \end{aligned}$$

Finally, from (8.11) we find the slope

$$\begin{aligned} B &= \frac{N\Sigma ml - \Sigma m \Sigma l}{\Delta} \\ &= \frac{5 \times 1622 - 30 \times 256.6}{200} = 2.06 \text{ cm/kg}. \end{aligned}$$

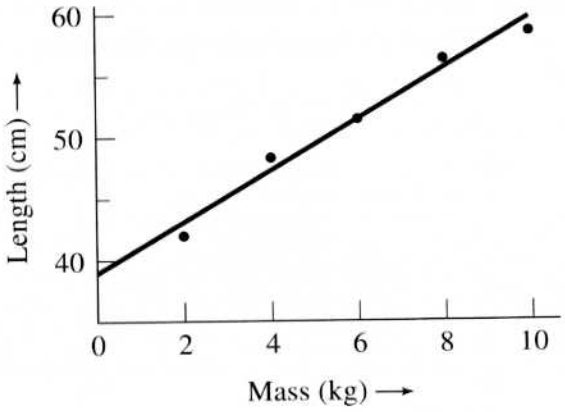

Figure 8.2. A plot of the data from Table 8.1 and the best-fit line (8.13).

A plot of the data and the line (8.13) using these values of A and B is shown in Figure 8.2. If the mass m stretches the spring to 53.2 cm, then according to (8.13) the mass is

$$m = \frac{l - A}{B} = \frac{(53.2 - 39.0)\text{ cm}}{2.06\text{ cm/kg}} = 6.9\text{ kg}.$$

Quick Check 8.1. Find the least-squares best-fit line $y = A + Bx$ through the three points $(x, y) = (-1, 0)$, $(0, 6)$, and $(1, 6)$. Using squared paper, plot the points and your line. [Note that because the three values of x (-1, 0, and 1) are symmetric about zero, $\Sigma x = 0$, which simplifies the calculation of A and B. In some experiments, the values of x can be arranged to be symmetrically spaced in this way, which saves some trouble.]

Now that we know how to find the best estimates for the constants A and B, we naturally ask for the uncertainties in these estimates. Before we can find these uncertainties, however, we must discuss the uncertainty σ_y in the original measurements of $y_1, y_2, \ldots, y_N$.

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, knowing how to calculate the uncertainty by analyzing the data themselves is important. 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 zero and the same width σ_y. This situation immediately suggests that a good estimate for σ_y would be given by a sum of squares with the familiar form

$$\sigma_y = \sqrt{\frac{1}{N}\Sigma(y_i - A - Bx_i)^2}. \tag{8.14}$$

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 (8.14). (See Problem 8.12.)

Unfortunately, as you may have suspected, the estimate (8.14) for σ_y is not quite the end of the story. The numbers A and B in (8.14) are the unknown true values of the constants A and B. In practice, these numbers 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 (8.14). 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

$$\sigma_y = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - A - Bx_i)^2}, \tag{8.15}$$

with A and B given by (8.10) and (8.11). If we already have an independent estimate of our uncertainty in $y_1, \ldots, y_N$, we would expect this estimate to compare with σ_y as computed from (8.15).

I will not attempt to justify the factor of $(N - 2)$ in (8.15) but 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 (8.14) and (8.15) are zero. Thus, the formula (8.14) (with $N = 2$ in the denominator) would give the absurd answer $\sigma_y = 0$; whereas (8.15), 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 (8.15) is reminiscent of the $(N - 1)$ that appeared in our estimate of the standard deviation of N measurements of one quantity x, in Equation (5.45). 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 computation 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. We can show (but will not do so here) that the number of degrees of freedom, *not* the number of measurements, is what should appear in the denominator of formulas such as (8.15) and (5.45). This fact explains why (8.15) contains the factor $(N - 2)$ and (5.45) 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 $y_1, \ldots, y_N$. I leave it as an exercise for you to check (Problem 8.16) that

$$\sigma_A = \sigma_y \sqrt{\frac{\Sigma x^2}{\Delta}} \tag{8.16}$$

and

$$\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}} \tag{8.17}$$

where Δ is given by (8.12) as usual.

The results of this and the previous two sections were based on the assumptions that the measurements of y were all equally uncertain and that any uncertainties in x were negligible. Although these assumptions often are justified, we need to discuss briefly what happens when they are not. First, if the uncertainties in y are not all equal, we can use the method of *weighted least squares,* as described in Problem 8.9. Second, if there are uncertainties in x but not in y, we can simply interchange the roles of x and y in our analysis. The remaining case is that in which both x and y have uncertainties—a case that certainly *can* occur. The least-squares fitting of a general curve when both x and y have uncertainties is rather complicated and even controversial. In the important special case of a *straight line* (which is all we have discussed so far), uncertainties in both x and y make surprisingly little difference, as we now discuss.

Suppose, first, that our measurements of x are subject to uncertainty but those of y are not, and we consider a particular measured point (x, y). This point and the true line $y = A + Bx$ are shown in Figure 8.3. The point (x, y) does not lie on the line because of the error—call it Δx—in our measurement of x. Now, we can see

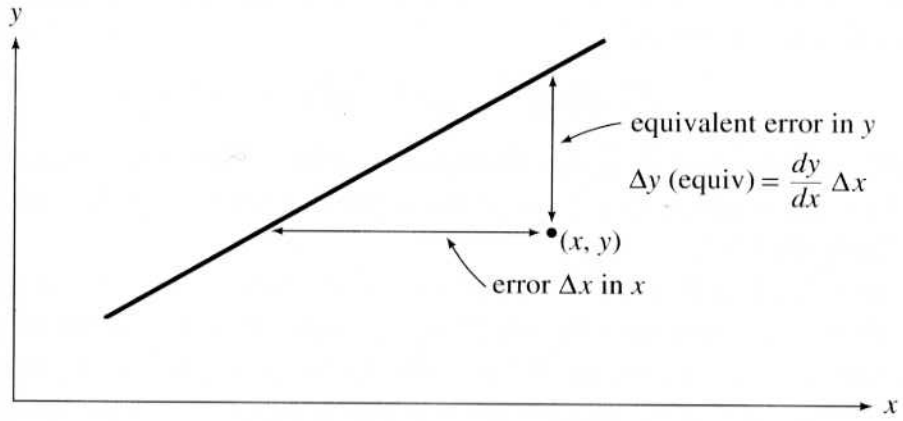

Figure 8.3. A measured point (x, y) and the line $y = A + Bx$ on which the point is supposed to lie. The error Δx in x, with y exact, produces the same effect as an error $\Delta y(\text{equiv}) = (dy/dx)\Delta x$ in y, with x exact. (Here, dy/dx denotes the slope of the expected line.)

easily from the picture that the error Δx in x, with no error in y, produces exactly the same effect as if there had been no error in x but an error in y given by

$$\Delta y(\text{equiv}) = \frac{dy}{dx}\Delta x \tag{8.18}$$

(where "equiv" stands for "equivalent"). The standard deviation σ_x is just the root-mean-square value of Δx that would result from repeating this measurement many times. Thus, according to (8.18), the problem with uncertainties σ_x in x can be replaced with an equivalent problem with uncertainties in y, given by

$$\sigma_y(\text{equiv}) = \frac{dy}{dx}\sigma_x. \tag{8.19}$$

The result (8.19) is true whatever the curve of y vs x, but (8.19) is especially simple if the curve is a straight line, because the slope dy/dx is just the constant B. Therefore, for a straight line

$$\sigma_y(\text{equiv}) = B\sigma_x. \tag{8.20}$$

In particular, if all the uncertainties σ_x are equal, the same is true of the equivalent uncertainties $\sigma_y(\text{equiv})$. Therefore, the problem of fitting a line to points (x_i, y_i) with equal uncertainties in x but no uncertainties in y is equivalent to the problem of equal uncertainties in y but none in x. This equivalence means we can safely use the method already described for either problem. [In practice, the points do not lie *exactly* on the line, and the two "equivalent" problems will not give *exactly* the same line. Nevertheless, the two lines should usually agree within the uncertainties given by (8.16) and (8.17). See Problem 8.17.]

We can now extend this argument to the case that *both* x and y have uncertainties. The uncertainty in x is equivalent to an uncertainty in y as given by (8.20). In addition, y is already subject to its own uncertainty σ_y. These two uncertainties are

independent and must be combined in quadrature. Thus, the original problem, with uncertainties in both x and y, can be replaced with an equivalent problem in which only y has uncertainty, given by

$$\sigma_y(\text{equiv}) = \sqrt{\sigma_y^2 + (B\sigma_x)^2}. \tag{8.21}$$

Provided all the uncertainties σ_x are the same, and likewise all the uncertainties σ_y, the equivalent uncertainties (8.21) are all the same, and we can safely use the formulas (8.10) through (8.17).

If the uncertainties in x (or in y) are not all the same, we can still use (8.21), but the resulting uncertainties will not all be the same, and we will need to use the method of weighted least squares. If the curve to which we are fitting our points is not a straight line, a further complication arises because the slope dy/dx is not a constant and we cannot replace (8.19) with (8.20). Nevertheless, we can still use (8.21) (with dy/dx in place of B) to replace the original problem (with uncertainties in both x and y) by an equivalent problem in which only y has uncertainties as given by (8.21).[1]

8.5 An Example

Here is a simple example of least-squares fitting to a straight line; it involves the constant-volume gas thermometer.

Example: Measurement of Absolute Zero with a Constant-Volume Gas Thermometer

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

$$T = A + BP. \tag{8.22}$$

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

$$A = -273.15°\text{C}$$

The constant B depends on the nature of the gas, its mass, and its volume.[2] 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 8.2. The student judged that his measurements of

[1] This procedure is quite complicated in practice. Before we can use (8.21) to find the uncertainty $\sigma_y(\text{equiv})$, we need to know the slope B (or, more generally, dy/dx), which is not known until we have solved the problem! Nevertheless, we can get a reasonable first approximation for the slope using the method of unweighted least squares, ignoring all of the complications discussed here. This method gives an approximate value for the slope B, which can then be used in (8.21) to give a reasonable approximation for $\sigma_y(\text{equiv})$.

[2] The difference $T - A$ is called the *absolute temperature.* Thus (8.22) can be rewritten to say that the absolute temperature is proportional to the pressure (at constant volume).

Table 8.2. Pressure (in mm of mercury) and temperature (°C) of a gas at constant volume.

Trial number i	"x" Pressure P_i	"y" Temperature T_i	$A + BP_i$
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

P had negligible uncertainty, and those of T were all equally uncertain with an uncertainty of "a few degrees." Assuming his points should fit a straight line of the form (8.22), he calculated his best estimate for the constant A (the absolute zero) and its uncertainty. What should have been his conclusions?

All we have to do here is use formulas (8.10) and (8.16), 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 ΣP, ΣP^2, ΣT, ΣPT. 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 8.2, we can calculate

$$\begin{aligned} \Sigma P &= 425, \\ \Sigma P^2 &= 37{,}125, \\ \Sigma T &= 260, \\ \Sigma PT &= 25{,}810, \\ \Delta &= N\Sigma P^2 - (\Sigma P)^2 = 5{,}000. \end{aligned}$$

In this kind of calculation, it is important to keep plenty of significant figures because we have to take differences of these large numbers. Armed with these sums, we can immediately calculate the best estimates for the constants A and B:

$$A = \frac{\Sigma P^2 \Sigma T - \Sigma P \Sigma PT}{\Delta} = -263.35$$

and

$$B = \frac{N \Sigma PT - \Sigma P \Sigma T}{\Delta} = 3.71.$$

This calculation already gives the student's best estimate for absolute zero, $A = -263$°C.

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 numbers are shown in the far right 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

$$\sigma_T = \sqrt{\frac{1}{N-2} \Sigma (T_i - A - BP_i)^2} = 6.7.$$

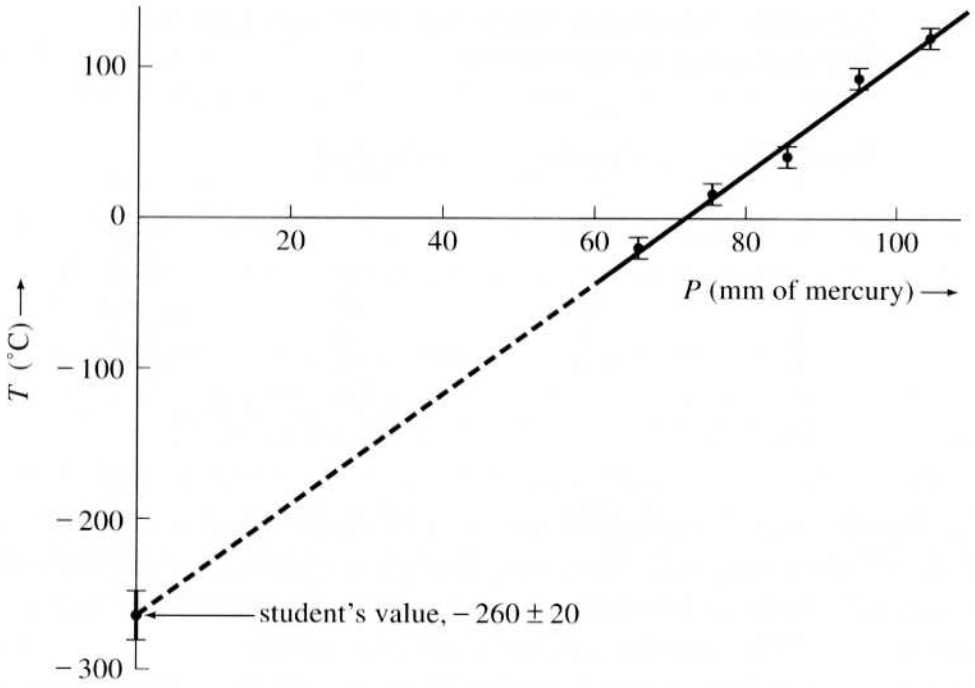

Figure 8.4. Graph of temperature T vs pressure P for a gas at constant volume. The error bars extend one standard deviation, σ_T, on each side of the five experimental points, and the line is the least-squares best fit. The absolute zero of temperature was found by extrapolating the line back to its intersection with the T axis.

This result 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 (8.16):

$$\sigma_A = \sigma_T\sqrt{\Sigma P^2/\Delta} = 18.$$

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

$$\text{absolute zero, } A = -260 \pm 20°\text{C},$$

which agrees satisfactorily with the accepted value, −273°C.

As is often true, these results become much clearer if we graph them, as in Figure 8.4. 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.

To find a value for absolute zero, the line was 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 magnification explains why the uncertainty in the value of absolute zero (±18°) is so much larger than that in the original temperature measurements (±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 section, I mention briefly a few more of these problems.

FITTING A POLYNOMIAL

Often, one variable, y, is expected to be expressible as a polynomial in a second variable, x,

$$y = A + Bx + Cx^2 + \cdots + Hx^n. \tag{8.23}$$

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, we can find best estimates for the constants $A, B, \ldots, H$ in (8.23) by an argument that exactly parallels that of Section 8.2, as I now outline.

To simplify matters, we suppose that the polynomial (8.23) is actually a quadratic,

$$y = A + Bx + Cx^2. \tag{8.24}$$

(You can easily extend the analysis to the general case if you wish.) 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 (8.24), 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 assumption lets us compute the probability of obtaining our observed values $y_1, \ldots, y_N$ in the familiar form

$$Prob(y_1, \ldots, y_N) \propto e^{-\chi^2/2}, \tag{8.25}$$

where now

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i - Cx_i^2)^2}{\sigma_y^2}. \tag{8.26}$$

[This equation corresponds to Equation (8.5) for the linear case.] The best estimates for A, B, and C are those values for which $Prob(y_1, \ldots, y_N)$ 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; see Problem 8.21):

$$\begin{aligned} AN + B\Sigma x + C\Sigma x^2 &= \Sigma y, \\ A\Sigma x + B\Sigma x^2 + C\Sigma x^3 &= \Sigma xy, \\ A\Sigma x^2 + B\Sigma x^3 + C\Sigma x^4 &= \Sigma x^2 y. \end{aligned} \tag{8.27}$$

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, and 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. (For an example, see Problem 8.22.)

The method of polynomial regression generalizes easily to a polynomial of any degree, although the resulting normal equations become 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, ... Unfortunately, the resulting normal equations that determine the best estimates for A, B, . . . can be difficult or impossible to solve. However, one large class of problems *can* always be solved, namely, those problems in which the function $y = f(x)$ depends linearly on the parameters A, B, ... These include all polynomials (obviously the polynomial (8.23) is linear in its coefficients A, B, . . .) but they also include many other functions. For example, in some problems y is expected to be a sum of trigonometric functions, such as

$$y = A \sin x + B \cos x. \tag{8.28}$$

For this function, and in fact for any function that is linear in the parameters A, B, . . . , the normal equations that determine the best estimates for A, B, . . . are simultaneous linear equations, which can always be solved (see Problems 8.23 and 8.24).

EXPONENTIAL FUNCTIONS

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

$$y = Ae^{Bx}, \tag{8.29}$$

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 $\lambda = 1/(RC)$, where R and C are the resistance and capacitance.

If the constants A and B in (8.29) are unknown, we naturally 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. We can, however, transform the nonlinear relation (8.29) 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 natural logarithm of (8.29) to give

$$\ln y = \ln A + Bx. \tag{8.30}$$

We see that, even though y is not linear in x, $\ln y$ *is*. This conversion of the nonlinear (8.29) into the linear (8.30) is useful in many contexts besides that of least-squares fitting. If we want to check the relation (8.29) 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 identified easily. (Such a plot is especially easy if you use "semilog" graph paper, on which the graduations on one axis are spaced logarithmically. Such paper lets you plot $\log y$ directly without even calculating it.)

The usefulness of the linear equation (8.30) 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 (8.30), or

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

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 (8.31). 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: A Population of Bacteria

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

$$N = N_0 e^{-t/\tau}, \tag{8.32}$$

where τ is called the population's *mean life* [closely related to the *half-life*, $t_{1/2}$; in fact, $t_{1/2} = (\ln 2)\tau$]. A biologist suspects that a population of bacteria is decreasing exponentially as in (8.32) and measures the population on three successive days; he obtains the results shown in the first two columns of Table 8.3. Given these data, what is his best estimate for the mean life τ?

Table 8.3. 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 (8.32), then the variable $z = \ln N$ should be linear in t:

$$z = \ln N = \ln N_0 - \frac{t}{\tau}. \tag{8.33}$$

Our biologist therefore calculates the three numbers $z_i = \ln N_i$ $(i = 0, 1, 2)$ shown in the third column of Table 8.3. Using these numbers, he makes a least-squares fit to the straight line (8.33) 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 coefficients 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

$$\sigma_z = \left|\frac{dz}{dy}\right| \sigma_y = \frac{\sigma_y}{y}. \tag{8.34}$$

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. The least-squares procedure can be modified to allow for different uncertainties in the measurements, provided the various uncertainties are known. (This method of *weighted least squares* is outlined in Problem 8.9). If we know that the measurements of $y_1, \ldots, y_N$ really are equally uncertain, then Equation (8.34) 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, we often cannot be sure that the uncertainties in $y_1, \ldots, y_N$ really are constant; so we can perhaps argue that we 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 which method is used makes little difference, as in the preceding example. In any event, when the uncertainties are unknown, 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. Many real problems, however, have more than two variables to be considered. For example, in studying the pressure P of a gas, we find that it depends on the volume V and temperature T, and we 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:

$$z = A + Bx + Cy. \tag{8.35}$$

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, and C are determined by normal equations of the form

$$\begin{aligned} AN + B\Sigma x + C\Sigma y &= \Sigma z, \\ A\Sigma x + B\Sigma x^2 + C\Sigma xy &= \Sigma xz, \\ A\Sigma y + B\Sigma xy + C\Sigma y^2 &= \Sigma yz. \end{aligned} \tag{8.36}$$

The equations can be solved for A, B, and C to give the best fit for the relation (8.35). This method is called *multiple regression* ("multiple" because there are more than two variables), but we will not discuss it further here.

Principal Definitions and Equations of Chapter 8

Throughout this chapter, we have considered N pairs of measurements $(x_1, y_1), \ldots, (x_N, y_N)$ of two variables x and y. The problem addressed was finding the best values of the parameters of the curve that a graph of y vs x is expected to fit. We assume that only the measurements of y suffered appreciable uncertainties, whereas those for x were negligible. [For the case in which both x and y have significant uncertainties, see the discussion following Equation (8.17).] Various possible curves can be analyzed, and there are two different assumptions about the uncertainties in y. Some of the more important cases are as follows:

A STRAIGHT LINE, $y = A + Bx$; EQUAL WEIGHTS

If y is expected to lie on a straight line $y = A + Bx$, and if the measurements of y all have the same uncertainties, then the best estimates for the constants A and B are:

$$A = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{\Delta}$$

and

$$B = \frac{N\Sigma xy - \Sigma x \Sigma y}{\Delta},$$

where the denominator, Δ, is

$$\Delta = N\Sigma x^2 - (\Sigma x)^2. \qquad \text{[See (8.10) to (8.12)]}$$

Based on the observed points, the best estimate for the uncertainty in the measurements of y is

$$\sigma_y = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - A - Bx_i)^2}. \qquad \text{[See (8.15)]}$$

The uncertainties in A and B are:

$$\sigma_A = \sigma_y \sqrt{\frac{\Sigma x^2}{\Delta}}$$

and

$$\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}}. \qquad \text{[See (8.16) \& (8.17)]}$$

STRAIGHT LINE THROUGH THE ORIGIN ($y = Bx$); EQUAL WEIGHTS

If y is expected to lie on a straight line through the origin, $y = Bx$, and if the measurements of y all have the same uncertainties, then the best estimate for the constant B is:

$$B = \frac{\Sigma xy}{\Sigma x^2}. \qquad \text{[See Problem 8.5]}$$

Based on the measured points, the best estimate for the uncertainty in the measurements of y is:

$$\sigma_y = \sqrt{\frac{1}{N-1}\Sigma(y_i - Bx_i)^2}$$

and the uncertainty in B is:

$$\sigma_B = \frac{\sigma_y}{\sqrt{\Sigma x^2}}. \qquad \text{[See Problem 8.18]}$$

WEIGHTED FIT FOR A STRAIGHT LINE, $y = A + Bx$

If y is expected to lie on a straight line $y = A + Bx$, and if the measured values y_i have different, known uncertainties σ_i, then we introduce the *weights* $w_i = 1/\sigma_i^2$, and the best estimates for the constants A and B are:

$$A = \frac{\Sigma wx^2 \, \Sigma wy - \Sigma wx \, \Sigma wxy}{\Delta}$$

and

$$B = \frac{\Sigma w \, \Sigma wxy - \Sigma wx \, \Sigma wy}{\Delta},$$

where

$$\Delta = \Sigma w \, \Sigma wx^2 - (\Sigma wx)^2. \qquad \text{[See Problem 8.9]}$$

The uncertainties in A and B are:

$$\sigma_A = \sqrt{\frac{\Sigma wx^2}{\Delta}}$$

and

$$\sigma_B = \sqrt{\frac{\Sigma w}{\Delta}}. \qquad \text{[See Problem 8.19]}$$

OTHER CURVES

If y is expected to be a polynomial in x, that is,

$$y = A + Bx + \ldots + Hx^n,$$

then an exactly analogous method of least-squares fitting can be used, although the equations are quite cumbersome if n is large. (For examples, see Problems 8.21 and 8.22.) Curves of the form

$$y = Af(x) + Bg(x) + \ldots + Hk(x),$$

where $f(x), \ldots, k(x)$ are known functions, can also be handled in the same way. (For examples, see Problems 8.23 and 8.24.)

If y is expected to be given by the exponential function

$$y = Ae^{Bx},$$

then we can "linearize" the problem by using the variable $z = \ln(y)$, which should satisfy the linear relation

$$z = \ln(y) = \ln(A) + Bx. \qquad \text{[See (8.31)]}$$

We can then apply the linear least-squares fit to z as a function of x. Note, however, that if the uncertainties in the measured values of y are all equal, the same is certainly *not* true of the values of z. Then, strictly speaking, the method of weighted least squares should be used. (See Problem 8.26 for an example.)

Problems for Chapter 8

For Section 8.2: Calculation of the Constants *A* and *B*

8.1. ★ Use the method of least squares to find the line $y = A + Bx$ that best fits the three points (1, 6), (3, 5), and (5, 1). Using squared paper, plot the three points and your line. Your calculator probably has a built-in function to calculate A and B; if you don't know how to use it, take a moment to learn and then check your own answers to this problem.

8.2. ★ Use the method of least squares to find the line $y = A + Bx$ that best fits the four points $(-3, 3)$, $(-1, 4)$, (1, 8), and (3, 9). Using squared paper, plot the four points and your line. Your calculator probably has a built-in function to calculate A and B; if you don't know how to use it, take a moment to learn and then check your own answers to this problem.