# 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

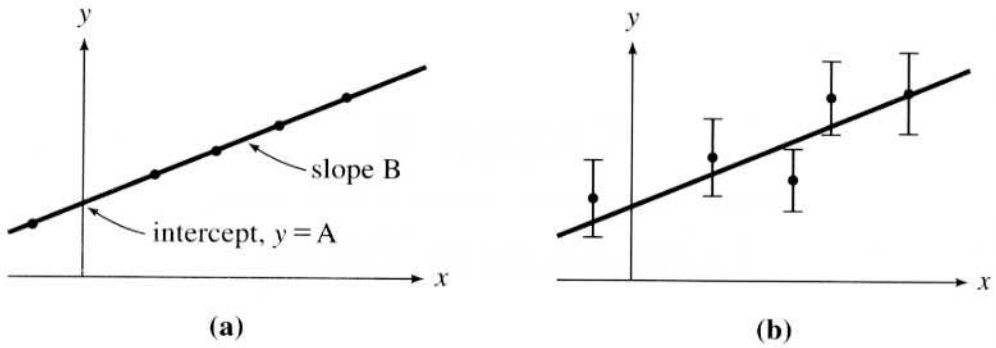


**Figure 8.1.** **(a)** If the two variables $y$ and $x$ are linearly related as in Equation (8.1), and if there were no experimental uncertainties, then the measured points $(x_i, y_i)$ would all lie exactly on the line $y = A + Bx$. **(b)** In practice, there always are uncertainties, which can be 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 $\sigma_y$ 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 $\sigma_y$. 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 $\chi^2$ in (8.5) is a minimum. (This is why the method is known as least-squares fitting.) To find these values, we differentiate $\chi^2$ 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 $\Sigma$. 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 $\Sigma xy$ 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 $\Sigma m_i$, $\Sigma l_i$, $\Sigma m_i^2$, 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 $\Delta$ 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

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