Parabola Volume 42, Issue 2 (2006)

Fitting Lines to Data

Bill McKee[1]

Introduction

We sometimes see in newspapers or on television situations where a straight line is drawn so as to approximately fit some data points. This can always be done by eye, using human judgment, but the results would then tend to vary depending on the person drawing the line. This article presents a rational way of constructing such a **line of best fit** and then goes on to generalise this to fitting other types of functions to data points.

In mechanics, Hooke's law states that the increase in length of a spring is proportional to the force applied to extend it. This is only an approximation applicable when the extension is relatively small. Suppose now that we were to perform an experiment in which we measured the extension of the spring (call it $\mathcal{E}$) for various values of the applied force (call this $\mathcal{F}$) and plotted the results on a piece of graph paper or on a computer screen. These will not in general lie exactly on a straight line and there are two main reasons for this. Firstly, Hooke's law is only an approximation to physical reality. Secondly, all experimental data are subject to errors. However, unless there is something horribly wrong with our experimental technique, we would expect that our data points would lie close to a straight line. We would want to find such a line, the slope of which is a measure of the constant of proportionality in Hooke's law.

Reminder - The Sigma Notation for Summation

If m is less than or equal to n (usually written as $m \leq n$) we use a shorthand notation to indicate sums of quantities as follows:

$$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + \ldots + a_n.$$

Often $m = 0$ or 1. Thus a polynomial of degree n can be written as

$$\sum_{i=0}^{n} b_i x^i = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n.$$

The use of this sigma notation is quite standard and saves writing long strings of symbols.

Linear Least-Squares Fitting

Rather than use $\mathcal{E}$ and $\mathcal{F}$, let us use x and y and suppose that we have n data points (x_i, y_i) for $i = 1, \ldots, n$. We now want to find a straight line $y = a + bx$ which best fits the data in some sense. The first thing to do is to decide what we mean by a good fit to the data. Suppose that we intended to use our line to estimate y for values of

[1]Dr Bill McKee is a Visiting Fellow at the School of Mathematics and Statistics, UNSW.

x which were not in the data (e.g. estimating $\mathcal{E}$ for values of $\mathcal{F}$ at which we did not have measurements). This is a common situation. We would hope that the difference between the true, but unknown, value of $\mathcal{E}$ and the value given by our predictive formula would be small so it makes sense to consider the differences between the values of y predicted by our straight line fit and the measured values y_i at each of our data points x_i. If we could make these small in some sense, we might expect that the errors in our formula would be small for other values of x too. Hence we are led to consider the **errors** defined by

$$E_i = y_i - a - bx_i.$$

There are n of these and they all depend on a and b. We want to choose a and b to make the E_i small in some sense. We could choose a and b to make our straight line pass exactly through two points but this would generally lead to large errors at the other data points. The values of a and b would also depend on which particular two points we chose. This is not a sensible strategy. Instead, it makes sense to consider the quantity Q defined by

$$Q = E_1^2 + E_2^2 + \ldots + E_n^2 = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \{y_i - a - bx_i\}^2 . \tag{1}$$

This is an overall measure of how good the approximation is at the data points. We now want to choose a and b to make it as small as possible. In passing, we should note that it makes no sense to consider $\sum_{i=1}^{n} E_i$ since large positive and negative errors could cancel out. We could try to minimise $\sum_{i=1}^{n} |E_i|$ but that leads to a much harder problem.

You will be familiar with the problem of minimising a function of one variable, say $f(x)$. We know that the minima occur where the derivative $f'(x)$ is zero and that there are tests for determining if such points are relative maxima, minima or neither. It is similar here. Q is a function of two variables a and b and the minimum will occur when the derivatives of Q with respect to a and b are both zero. So, differentiating Q with respect to a and setting the result to zero gives

$$\sum_{i=1}^{n} (-2)(y_i - a - bx_i) = 0. \tag{2}$$

Similarly, differentiating Q with respect to b and setting the result to zero gives

$$\sum_{i=1}^{n} (-2x_i)(y_i - a - bx_i) = 0. \tag{3}$$

Re-arranging these and noting that $\sum_{i=1}^{n} a = na$, leads to what are known in the trade as the **normal equations**:

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{4}$$

$$\text{and} \qquad a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{5}$$

This is a simple system of two simultaneous linear equations for a and b. The coefficients are all known since the x_i and y_i are the known data points. You should all be able to show that the solution is

$$a = \Delta\{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i y_i)(\sum_{i=1}^{n} x_i)\} \quad (6)$$

$$\text{and} \qquad b = \Delta\{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)\} \quad (7)$$

where

$$\Delta = \frac{1}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} \, .$$

In passing, let us consider what would happen if we were to set $b = 0$ in (1). We would then be trying to approximate our data by a constant. Q would then be a function only of a and only (2) with $b = 0$ therein is relevant. This has the solution:

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i$$

i.e. the best least-squares approximation of the data by a constant is just the average of all the y values. This seems eminently reasonable.

We also need to convince ourselves that we have, in fact, minimised Q. There are tests for this which are generalisations of the second derivative test for functions of one variable. These are the material of second-year University courses and are too complicated to discuss here. It should be rather obvious that solving the normal equations will give us a minimum rather than a maximum since we can make Q as large as we like by choosing very large values of a to make the line $y = a + bx$ pass nowhere near any of the data points.

The process we have just described of fitting a straight line to data by minimising Q is called **linear regression** by statisticians. It should be noted that our line of best fit will not, in general, pass through any of our data points. Thus, in our Hooke's law example, we know that $\mathcal{E} = 0$ when $\mathcal{F} = 0$ yet our line of best fit need not pass exactly through this point.

OK, so far, so good. Let us see how this works in a specific example. Consider the following data:

x_i	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0
y_i	1.3	3.5	4.2	5.0	7.0	8.8	10.1	12.5	13.0	15.6	16.1

Feeding these numbers into (6) and (7) gives the best least-squares fit to the data as

$$y = -.2763636364 + 1.517272727x \quad (8)$$

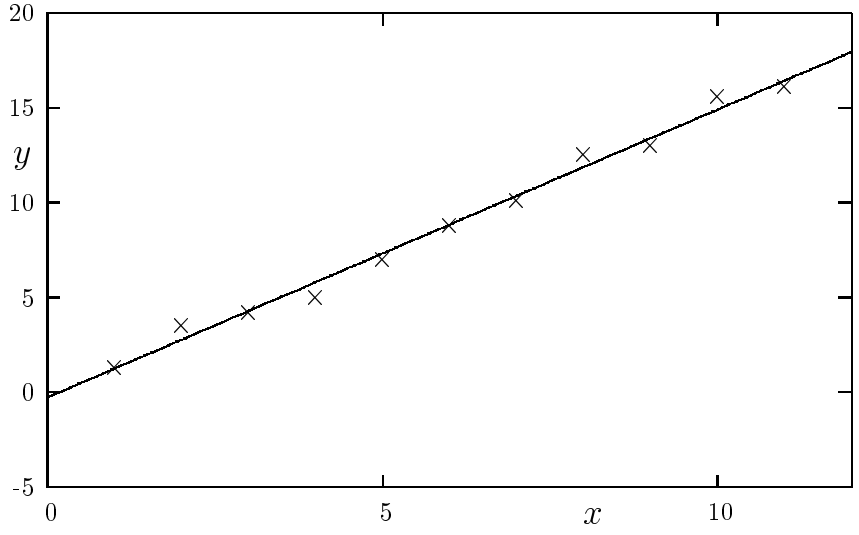

Figure 1: The crosses represent the data shown in the table and the straight line is the least-squares fit obtained by minimising Q as defined by equation (1).

This is shown in Figure 1 and the fit is seen to be quite good in this case. Going back to our mechanical example, supposing that $x = \mathcal{F}$ and $y = \mathcal{E}$ we could use (8) to predict $\mathcal{E}$ given $\mathcal{F}$. On the other hand, we might want to predict $\mathcal{F}$ given $\mathcal{E}$. One way of doing this would be to solve (8) for x as a function of y to give

$$x = (y + 0.2763636364)/1.517272727 = .6590772919y + .1821449971 \tag{9}$$

An alternative strategy would be to reverse the roles of x and y from the beginning and so seek a line of the form

$$x = \alpha + \beta y$$

where α and β are chosen to minimise

$$W = \sum_{i=1}^{n} \{x_i - \alpha - \beta y_i\}^2 . \tag{10}$$

It is not necessary to work this all out again. All we have to do is swap the x_i and y_i in (6) and (7) and replace a by α and b by β. Doing this, we find the best least-squares fit as

$$x = .2387715344 + .6526623390y \tag{11}$$

which is different from (9). On reflection, this is not at all surprising since the two equations were obtained by minimising the different quantities Q and W. In this particular example, the two lines are remarkably similar. In fact, we have not plotted the line given by (11) on Figure 1 since it is so close to that given by (8) that it is hard for the eye to distinguish them. The basic reason for this similarity is that the given data do, in fact, lie quite close to a straight line. If we now consider another data set for which the data points lie not so close to a straight line, things are somewhat different, as shown in Figure 2. In this example, the x_i are measurements of my systolic blood pressure and the y_i are measurements of my diastolic blood pressure. There is a general tendency for one to be high or low when the other is high or low but the relationship between the two is not nearly as marked as for the first data set and the data are more scattered.

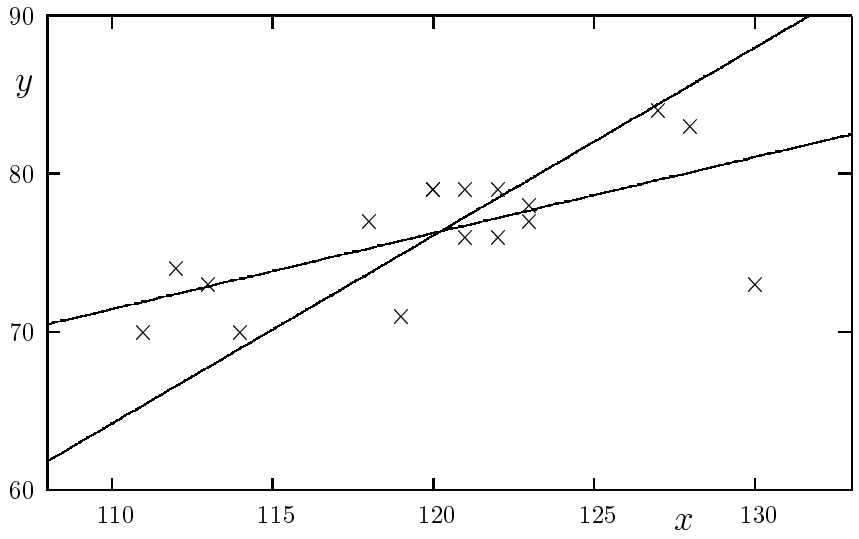

Figure 2: Least-squares fitting of blood pressure data.The line of greater slope minimises W given by (10) and the line of smaller slope minimises Q given by (1).

Fitting other types of functions

For any given data set of (x_i, y_i) values, we can always fit a straight line to the data as described above. However, this may not be particularly sensible. For example, suppose we had meteorological records of the atmospheric temperature measured at Sydney airport every hour for one week. Although there will be considerable variability in the data, we know that temperatures tend to follow a daily cycle and so fitting the data with sines and cosines with a period of one day would probably be more

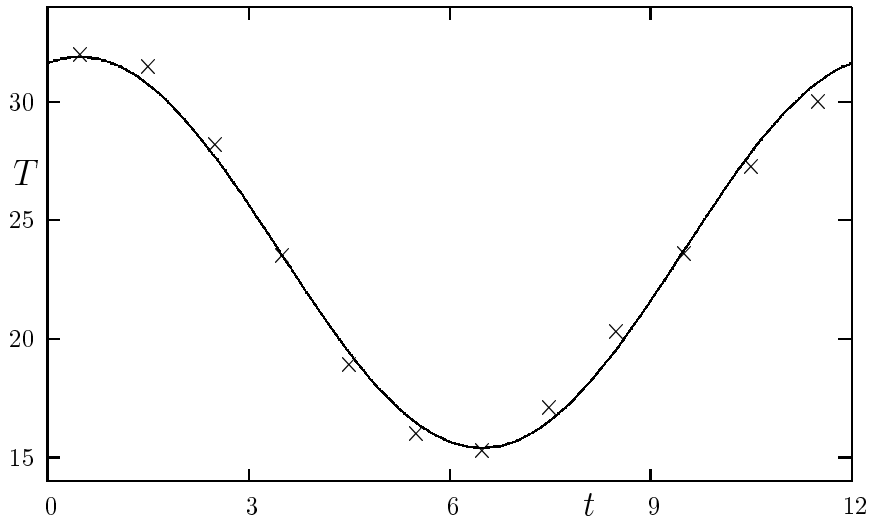

Figure 3: The crosses give the average daily maximum temperature at Mildura airport for each month (plotted at the middle of each month) and the solid line is the least-squares fit obtained using equation (14).

sensible. Tidal observations are another case where we know something about the nature of the data before we start. Indeed, tidal predictions are basically made this way. There are a great many periodic components involved due to the complicated motions of the earth, moon and sun but the periods of these are known very accurately from astronomical theory and observations. Past tidal measurements at any given port are then fitted with a large number of sines and cosines with these known periods and the results are used to predict tides at that port into the future.

To see how this would be done, suppose that we wished to use our least-squares method to fit a given data set of n data points (x_i, y_i) for $i = 1, \ldots, n$, with a function of the form

$$y = \sum_{j=1}^{M} A_j \Phi_j(x) \tag{12}$$

where the $\Phi_j(x)$ are appropriately-chosen functions. If we were trying to fit a polynomial of degree $M-1$ to the data we would take $\Phi_j(x) = x^{j-1}$ for $j = 1, \ldots, M$. Using $M = 2$ with $\Phi_1(x) = 1$, $\Phi_2(x) = x$, $A_1 = a$ and $A_2 = b$ gives us the straight-line fitting considered above. For tidal data, the $\Phi_j(x)$ would be sines and cosines with the periods

known to be relevant to tides. We would now choose the coefficients A_j to minimise

$$R = \sum_{i=1}^{n} \{y_i - \sum_{j=1}^{M} A_j \Phi_j(x_i)\}^2. \tag{13}$$

This is just a generalisation of the problem considered earlier as R is a function of the M variables A_k for $k = 1, \ldots, M$. The minimum will occur where the derivative of R with respect to A_k vanishes for $k = 1, \ldots, M$. This leads to a system of M simultaneous linear equations to solve for the M coefficients A_k. This is a straightforward task although the details will not be gone into here. Once this is done, we construct our approximation from (12). An example is shown in Figure 3 in which the data points represent the average daily maximum temperature at Mildura airport for each month (plotted at the middle of each month and obtained from the Bureau of Meteorology website at http://www.bom.gov.au). The unit of time t employed is one month and the unit of temperature T is one degree Celsius. The curve is a least-squares fit of the form

$$y = A_1 + A_2 \cos \frac{\pi t}{6} + A_3 \sin \frac{\pi t}{6} \tag{14}$$

which is periodic in t with period 12 months. The fit is seen to be quite good.

Postscript

Least-squares fitting leads to relatively easy equations to solve for the coefficients in (12) but is not the only way of fitting curves to data. This is an important area in practical applications and still a topic of active research.

Exercise

If we define the average values of the x_i and y_i to be

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{and} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

show that the straight line fit $y = a + bx$ with a and b determined by (6) and (7) can be written as

$$y - \bar{y} = b(x - \bar{x}) \, .$$

Hence our least-squares straight line fit passes through the point whose coordinates are the average values of the x_i and y_i.