APPENDIX N
Fitting Functions to Data

Frequently, you would like to estimate the best values of the coefficients in an equation from experimental data. The equation might be a theoretical law or just a polynomial, but the procedure is the same. Let y be the dependent variable in the equation, b_i be the coefficients in the equation, and x_i be the independent variables in the equation so that the model is of the form

$$y = f(b_0, b_1, \ldots ; x_0, x_1, \ldots) \tag{N.1}$$

Let ϵ represent the error between the observation of y, Y, and the predicted value of y using the values of x_i and the estimated values of the coefficients b_i:

$$Y = y + \epsilon \tag{N.2}$$

The classical way to get the best estimates of the coefficients is by least squares, that is, by minimizing the sum of the squares of the errors (of the deviations) between Y and y for all the j sets of experimental data, $j = 1$ to p:

$$\text{Minimize } F = \sum_{j=1}^{p} (\epsilon_j)^2 = \sum_{j=1}^{p} (Y_j - f_j)^2 \tag{N.3}$$

Let us use a model linear in the coefficients with one independent variable x

$$y = b_0 + b_1 x \tag{N.4}$$

(in which x_0 associated with b_0 always equals 1 in order to have an intercept) to illustrate the principal features of the least-squares method to estimate the model coefficients. The objective function is

$$F = \sum_{j=1}^{p} (Y_j - y_j)^2 = \sum_{j=1}^{p} (Y_j - b_0 - b_1 x_j)^2 \tag{N.5}$$

There are two unknown coefficients, b_0 and b_1, and p known pairs of experimental values of Y_j and x_j. We want to minimize F with respect to b_0 and b_1. Recall from calculus that you take the first partial derivatives of F and equate them to zero to get the necessary conditions for a minimum.

$$\frac{\partial F}{\partial b_0} = 0 = 2\sum_{j=1}^{p}(Y_j - b_0 - b_1 x_j)(-1) \tag{N.6a}$$

$$\frac{\partial F}{\partial b_1} = 0 = 2\sum_{j=1}^{p}(Y_j - b_0 - b_1 x_j)(-x_j) \tag{N.6b}$$

Rearrangement yields a set of linear equations in two unknowns, b_0 and b_1:

$$\sum_{j=1}^{p} b_0 + \sum_{j=1}^{p} b_1 x_j = \sum_{j=1}^{p} Y_j$$

$$\sum_{j=1}^{p} b_0 x_j + \sum_{j=1}^{p} b_1 x_j^2 = \sum_{j=1}^{p} x_j Y_j$$

The summation $\Sigma_{j=1}^{p}\, b_0$ is $(p)(b_0)$ and in the other summations the constants b_0 and b_1 can be removed from within the summation signs so that

$$b_0(p) + b_1\sum_{j=1}^{p} x_j = \sum_{j=1}^{p} Y_j \tag{N.7a}$$

$$b_0\sum_{j=1}^{p} x_j + b_1\sum_{j=1}^{p} x_j^2 = \sum_{j=1}^{p} x_j Y_j \tag{N.7b}$$

The two linear equations above in two unknowns, b_0 and b_1, can be solved quite easily for b_0 the intercept and b_1 the slope.

Example N.1 Application of Least Squares

Fit the model $y = \beta_0 + \beta_1 x$ to the following data (Y is the measured response and x the independent variable).

x	Y
0	0
1	2
2	4
3	6
4	8
5	10

(*Continues*)

AppN.qxd 3/10/12 9:26 PM Page 926

Example N.1 Application of Least Squares (*Continued*)

Solution

The computations needed to solve Equation (N.7) are

$$\sum x_j = 15 \qquad \sum x_j Y_j = 110$$

$$\sum Y_j = 30 \qquad \sum x_j^2 = 55$$

Then

$$6b_0 + 15b_1 = 30$$
$$15b_0 + 55b_1 = 110$$

Solution of these two equations yields

$$b_0 = 0 \qquad b_1 = 2$$

and the model becomes $\hat{Y} = 2x$, where $\hat{Y}$ is the predicted value for a given x.

The least-squares procedure outlined above can be extended to any number of variables as long as the model is linear in the coefficients. For example, a polynomial

$$y = a + bx + cx^2$$

is linear in the coefficients (but not in x), and would be represented as

$$y = b_0 + b_1 x_1 + b_2 x_2$$

where $a = b_0$, $b = b_1$, $c = b_2$, $x_1 = x$, and $x_2 = x^2$. Linear equations equivalent to Equation (N.7) with several independent variables can be solved via a computer. If the equation you want to fit is nonlinear in the coefficients, such as

$$y = b_0 e^{b_1 x} + b_2 x^2$$

you can minimize F in Equation (N.3) directly by the computer program on the disk in the back of this book, or by using a nonlinear least-squares computer code taken from a library of computer codes.

Additional useful information can be extracted from a least-squares analysis if four basic assumptions are made in addition to the presumed linearity of y in x:

1. The x's are deterministic variables (not random).
2. The variance of ϵ_j is constant or varies only with the x's.
3. The observations Y_j are mutually statistically independent.
4. The distribution of Y_j about y_j given x_j is a normal distribution.

For further details, see Box, Hunter, and Hunter[1] or Box and Draper.[2]

[1]G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. New York: Wiley-Interscience (1978).

[2]G. E. P. Box and N. R. Draper. *Empirical Model Building and Response Surfaces*. New York: Wiley (1987).