

11 More Approximation of Functions

The previous forward, backward, and central difference methods, and the Lagrange interpolation method, are all designed to allow one to approximate functions under the assumption that the data are exact or nearly exact.

If the data are assumed to be noisy, then the approximation of the data with a function is usually done with a *regression* method, of which the most commonly used method is that of a linear least squares approximation.

11.1 Linear Least Squares

Let us assume that we have data points

$$(x_1, y_1), \quad (x_2, y_2), \quad \dots, \quad (x_n, y_n),$$

and that these points lie along a more or less straight line

$$Y = mX + b$$

How do we find the values of m and b that provide a “best” fit of the straight line to the data?

11.2 Linear least squares

Consider the discussion on pages 237ff. in the text. We will assume that we are looking for m and b . If we had values for m and b , then for a given x_i , the approximate y -value would be $mx_i + b$. We will define the “best” straight line as the one for which the sums of the squares of the vertical distances from y_i to $mx_i + b$ is minimised.

This sets up a calculus problem: find the values of m and b such that

$$S = \sum_{i=1}^n (y_i - (mx_i + b))^2$$

is minimised.

This is a two-variable calculus problem. Compute the first partial derivatives

$$\frac{\partial S}{\partial m} = \sum_{i=1}^n 2(y_i - (mx_i + b))(-x_i) = 2 \sum_{i=1}^n (-x_i y_i + mx_i^2 + bx_i)$$

and

$$\frac{\partial S}{\partial b} = \sum_{i=1}^n 2(y_i - (mx_i + b))(-1) = 2 \sum_{i=1}^n (-y_i + mx_i + b)$$

then set both of these equal to zero, and solve for m and b . This gives us

$$2 \sum_{i=1}^n (-x_i y_i + m x_i^2 + b x_i) = 0$$

$$2 \sum_{i=1}^n (-y_i + m x_i + b) = 0$$

which then becomes

$$m \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i$$

$$m \sum_{i=1}^n x_i + b \sum_{i=1}^n 1 = \sum_{i=1}^n y_i$$

We solve

$$m \left(\sum_{i=1}^n x_i^2 \sum_{i=1}^n x_i \right) + b \left(\sum_{i=1}^n x_i \right)^2 = \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i$$

$$m \left(\sum_{i=1}^n x_i^2 \sum_{i=1}^n x_i \right) + n b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i \sum_{i=1}^n x_i^2$$

to get

$$b = \frac{\left(\sum_{i=1}^n x_i y_i \right) \left(\sum_{i=1}^n x_i \right) - \left(\sum_{i=1}^n y_i \right) \left(\sum_{i=1}^n x_i^2 \right)}{\left(\sum_{i=1}^n x_i \right)^2 - n \left(\sum_{i=1}^n x_i^2 \right)}$$

and

$$mn \left(\sum_{i=1}^n x_i^2 \right) + nb \left(\sum_{i=1}^n x_i \right) = n \sum_{i=1}^n x_i y_i$$
$$m \left(\sum_{i=1}^n x_i \right)^2 + nb \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \sum_{i=1}^n x_i$$

to get

$$m = \frac{n \left(\sum_{i=1}^n x_i y_i \right) - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2}$$

These solutions,

$$m = \frac{n \left(\sum_{i=1}^n x_i y_i \right) - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2}$$
$$b = \frac{\left(\sum_{i=1}^n x_i y_i \right) \left(\sum_{i=1}^n x_i \right) - \left(\sum_{i=1}^n y_i \right) \left(\sum_{i=1}^n x_i^2 \right)}{\left(\sum_{i=1}^n x_i \right)^2 - n \left(\sum_{i=1}^n x_i^2 \right)}$$

are the linear least squares regression of y upon x for the given data.

11.3 Data not linear?

What do we do if the data aren't linear?

Option one is to redo the whole problem from start to finish. If, instead of a linear function $Y = mX + b$, we expect a quadratic function $Y = aX^2 + bX + c$, then we should be minimising

$$Q = \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))^2$$

by setting $\frac{\partial Q}{\partial a}$, $\frac{\partial Q}{\partial b}$, and $\frac{\partial Q}{\partial c}$ all equal to zero, and then solving. This would give us three equations in three unknowns a , b , and c , and in theory we could solve.

Option two, which works in many instances, is to massage the data. If we expect that the function is exponential, $Y = Ae^{BX}$ for constants A and B , then we have $\log Y = \log A + BX$, which has suddenly become a linear equation with coefficients B and $\log A$ to be found by linear least squares applied to the points $(x_i, \log y_i)$.