The Newton-Raphson Method

The Newton–Raphson method for minimising a function $S = S(\theta)$ is based on the following quadratic approximation to S:

$$S_q(\theta) = S(\theta_r) + \gamma_r'(\theta - \theta_r) + \frac{1}{2}(\theta - \theta_r)'H_r(\theta - \theta_r). \tag{1}$$

Here $\gamma_r = \partial S(\theta_r)/\partial\theta$ stands for the derivative of the function evaluated at θ_r, whilst $H_r = \partial\{\partial S(\theta_r)/\partial\theta\}'/\partial\theta$ is the Hessian matrix comprising the scond-order partial derivatives of the function, also evaluated at θ_r. By differentiating S_q in respect of θ and setting the result to zero, we obtain the condition

$$0 = \gamma_r' + (\theta - \theta_r)'H_r. \tag{2}$$

The value which minimises the function is therefore

$$\theta_{r+1} = \theta_r - H_r^{-1}\gamma_r; \tag{3}$$

and this expression describes the $(r + 1)$th iteration of the Newton–Raphson algorithm. If the function to be minimised is indeed a concave quadratic, then the Newton–Raphson procedure will attain the minimum in a single step. Notice also that, if $H = I$, then the method coincides with the method of steepest descent. In the case of $H = I$, the contours of the quadratic function are circular.

The disadvantages of the Newton–Raphson procedure arise when the value of the Hessian matrix at θ_r is not positive definite. In that case, the step from θ_r to θ_{r+1} is liable to be in a direction which is away from the minimum value. This hazard can be illustrated by a simple diagram which relates to the problem of finding the minimum of a function defined over the real line. The problems only arise when the approximation θ_r is remote from the true minimum of the function. Of course, in the neighbourhood the minimising value, the function is concave; and, provided that the initial approximation θ_0, with which the iterations begin, is within this neighbourhood, the Newton–Raphson procedure is likely to perform well.

The Minimisation of a Sum of Squares

In statistics, we often encounter the kind of optimisation problem which requires us to minimise a sum-of-squares function

$$S(\theta) = \varepsilon'(\theta)\varepsilon(\theta), \tag{4}$$

wherein $\varepsilon(\theta)$ is a vector of residuals which is a nonlinear function of a vector θ. The value of θ corresponding to the minimum of the function commonly represents the least-squares estimate of the parameters of a statistical model. Such problems may be approached using the Newton–Raphson method which we described in the previous section. However, the specialised nature of the function $S(\theta)$ allow us to pursue a method which avoids the trouble of finding its second-order derivatives and which has other advantages as well. This is the Gauss–Newton method, and it depends upon using a linear approximation of the function $\varepsilon = \varepsilon(\theta)$. In the neighbourhood of θ_r, the approximating function is

$$e = \varepsilon(\theta_r) + \frac{\partial \varepsilon(\theta_r)}{\partial \theta}(\theta - \theta_r), \tag{5}$$

where $\partial\varepsilon(\theta_r)/\partial\theta$ stands for the first derivative of $\varepsilon(\theta)$ evaluated at $\theta = \theta_r$. This gives rise, in turn, to an approximation to S in the form of

$$\begin{aligned} S_g = \varepsilon'(\theta_r)\varepsilon(\theta_r) &+ (\theta - \theta_r)'\left\{\frac{\partial \varepsilon(\theta_r)}{\partial \theta}\right\}'\left\{\frac{\partial \varepsilon(\theta_r)}{\partial \theta}\right\}(\theta - \theta_r) \\ &+ 2\varepsilon'(\theta_r)\frac{\partial \varepsilon(\theta_r)}{\partial \theta}(\theta - \theta_r). \end{aligned} \tag{6}$$

By differentiating S_g in respect of θ and setting the result to zero, we obtain the condition

$$0 = 2(\theta - \theta_r)'\left\{\frac{\partial \varepsilon(\theta_r)}{\partial \theta}\right\}'\left\{\frac{\partial \varepsilon(\theta_r)}{\partial \theta}\right\} + 2\varepsilon'(\theta_r)\frac{\partial \varepsilon(\theta_r)}{\partial \theta}. \tag{7}$$

The value which minimises the function S_g is therefore

$$\theta_{r+1} = \theta_r - \left[\left\{\frac{\partial \varepsilon(\theta_r)}{\partial \theta}\right\}'\left\{\frac{\partial \varepsilon(\theta_r)}{\partial \theta}\right\}\right]^{-1}\left\{\frac{\partial \varepsilon(\theta_r)}{\partial \theta}\right\}'\varepsilon(\theta_r). \tag{8}$$

This equation represents the algorithm of the Gauss–Newton procedure, and it provides the formula by which we can find the $(r+1)$th approximation to the value which minimises sum of squares once we have the rth approximation.

The affinity of the Gauss–Newton and the Newton–Raphson methods is confirmed when we recognise that the matrix in (8) is simply an approximation to the Hessian matrix of the sum-of-squares function which is

$$\frac{\partial(\partial S/\partial\theta)'}{\partial \theta} = 2\left[\left(\frac{\partial \varepsilon}{\partial \theta}\right)'\left(\frac{\partial \varepsilon}{\partial \theta}\right) + \sum_t \varepsilon_t\left\{\frac{\partial(\partial\varepsilon_t/\partial\theta)'}{\partial \theta}\right\}'\right]. \tag{9}$$

The matrix of the Gauss–Newton procedure is always positive semi-definite; and, in this respect, the procedure has an advantage over the Newton–Raphson procedure.