The Use of NAG Optimisation Routines for Parameter Estimation

Geoff Morgan, NAG Ltd

The NAG Fortran Library contains an increasing number of routines to fit statistical models. However, for more specialist examples the user has to use the optimisation routines in the NAG E04 chapter. The statistical model will give the probability of observing a data value in terms of a set of parameters. The values of the parameters have to be estimated from the data in such a way that the model best fits the data. For example if we assume that the observations have a Poisson distribution then the probability of observing the value y is

$$P(Y=y) = \frac{m^{y} e^{-m}}{y!}$$

where the parameter $\mathbf{m}$ is the mean of the distribution. The mean may be related to another observed variable, say x, through the model:

$$m = e^{b_0+b_1x}$$

Given a number of observations of y and x we need to find estimates of the parameters b_0 and b_1. Such a model could be fitted using the NAG routine for generalised linear models with Poisson errors, G02GCF. However, if the model were

$$m = a + e^{b_0+b_1x}$$

the user would have to fit the model using an E04 routine.

The criterion often used to find estimators of the parameters of a statistical model is maximum-likelihood. In maximum-likelihood estimation the parameters are chosen so that the observed data are the most likely outcome. To find such estimators first the likelihood function has to be found. The likelihood is the probability distribution for the sample considered as a function of the parameters rather than as a function of the random variable. In the case of a sample of independent observations the likelihood is the product of the probability distribution for each observation. For example, for the Poisson distribution the likelihood is

$$\text{likelihood} = \prod_i \frac{m_i^{y_i} e^{-m_i}}{y_i!}$$

In many cases the log-likelihood is a simpler function, in the Poisson example

$$\text{log-likelihood} = \text{const} + \sum_i y_i \log(m_i) - \sum_i m_i$$

where **const** is independent of the parameters. In the case of an independent sample from the Normal distribution, the log-likelihood is a sum of squares function and so the maximum-likelihood estimators are the same as the least-squares estimators.

Having found the likelihood or log-likelihood function it can be maximised using an appropriate NAG E04 routine to obtain the parameter estimations. In the case of the Normal distribution there are specialist non-linear least-squares routines which should be used.

In addition to the parameter estimates the user will generally require the variance-covariance matrix of the parameter estimates. These can be used to test hypotheses about the parameters or to set confidence intervals for the parameters. It is also advisable to examine the correlation between the parameter estimates. High correlations indicate that either a different parameterisation should be considered or, if that is not possible, inferences should only be made jointly for the correlated parameters. For maximum-likelihood estimators that have been found at an interior maximum, it can be shown that for large samples the variance-covariance matrix can be approximated by minus the inverse of the expected value of the matrix of second derivatives of the log-likelihood. However, as in many cases the expected values of the derivatives are difficult to compute and the observed values are used instead, ie, minus the inverse of the Hessian matrix.

Many of the NAG E04 routines return approximations to the Hessian matrix, or to the Cholesky factors of the Hessian matrix and many users are misled into attempting to use this information for the variance-covariance calculations. In fact the information returned is fine in the context of restarting the optimisation process but is generally not good enough for the statistical calculation. This is particularly so when the routine returns a non-zero value for IFAIL, such as IFAIL=3 for E04WDF, in which case the routine usually returns with the unit matrix as an approximation to the Hessian.

Instead the user is recommended to use E04XAF which has been included in the Library to give a more reliable means of estimating the Hessian matrix. The user should also note that E04YCF has been included to provide estimates of the variance-covariance matrix following the use of one of the unconstrained non-linear least squares routines.

Tel: +44 (0)1865 511245
Email: sales@nag.co.uk