Approximate Confidence Intervals

REFERENCE: IS Chan, AA Goldstein, and JB Bassingthwaighte, SENSOP: A Derivative-Free Solver for nonlinear least squares with sensitivity scaling, ABME, Vol 21, pp. 621-631, 1993. (Our files ARticle #387).

NOTA BENE: “Near the solution for small residuals” the calculation of the covariance matrix is an approximation. (2nd column, top of page 625).

Let the data points being fit be given as

$$\{y(i), i = 1, nh\}.$$

Let the associated weights be given as

$$\{w(i), i = 1, nh\}.$$

Let the fitting parameters be given as

$$\{x(j), j = 1, nx\}.$$

Let

$f((x(j), j = 1, nx), i) = h(i)$ be the model fit to the data which minimizes the sum of the squares of the weighted residuals (SSR).

$$SSR = \sum_{i=1}^{nh} (h(i) - y(i))^2 \cdot w(i).$$

Define the following matrices,

$W = \begin{bmatrix} w_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & w_{nh} \end{bmatrix}$, the matrix of weights, and

$S = \begin{bmatrix} s_{ik} \end{bmatrix}$, a matrix with dimensions nh by nx, where s_{ik} is defined as follows.

For each parameter x(k), k=1 to nx
    Perturb x(k) by
    x(k)=x(k)-delx(k)
    For each data point i, i=1 to nh
        Calculate the model value using the optimized values for (x(j), $j \neq k$)

and the perturbed value x(k) giving h(i,k).

$$s_{ik} = \left(\frac{h_i - h_{ik}}{delx_k} . \right)$$

End loop over each data point
Restore x(k) by to its optimized value
x(k)=x(k)+delx(k).
End loop over each parameter.

We define a final matrix, Cov(X), the covariance matrix (with dimensions nx by nx) by

$$Cov(X) = \frac{SSR}{nh - nx} \cdot [S^T \cdot W \cdot S]^{-1} , \quad \textbf{EQUATION 18}$$

where $[S^T \cdot W \cdot S]^{-1}$ is the inverse of the product of the transpose of S multiplying W multiplying S.

Finally the 95% confidence limit for x(j) is given by

$$\pm t_{nh - nx, 0.975} \cdot (Cov[jj])^{1/2} , \quad \textbf{EQUATION 19}$$

where $t_{nh - nx, 0.975}$ is the Student's t-distribution with nh-nx degrees of freedom and $Cov[jj]$ is the j'th diagonal element of Cov(X).

"CONFIDENCE INTERVALS OF ESTIMATED PARAMETERS

The covariance matrix of the solution parameters can be estimated by the Hessian matrix at the solution (i.e., the second derivative matrix). Near the solution for small residuals, the right hand side of eq. 18 (above) is an approximation to the covariance matrix. Based on the approximation, the 95% confidence interval for parameter x(j) is given by eq 19.

(CAVEATS): "The estimates of the confidence interval will be underestimated if the model function is highly nonlinear and the residuals are large. To measure the goodness of estimated confidence interval, one may perform a simulation study by repeated trials of fitting model solutions to data to which the experimentally appropriate levels of random noise have been added. If high accuracy of the confidence interval is desired, one may try a more computationally intensive method such as that of Duncan (An empirical study of jackknife-constructed confidence regions in non-linear regression. Technometrics 20:123-129; 1978) in which asymmetry of the upper and lower limits is properly treated. An extensive comparison of various methods for the confidence intervals was reported by Donaldson and Schnabel (Computational experience with confidence regions and confidence intervals for nonlinear least squares. Technometrics 29:67-82; 1987)."

Validation of calculations:

We took the simple equation ***y=t-2*** from t=0 to 3.0 in increments 0f 0.1 and added noise to it which was uniformly distributed from -0.5 to 0.5. Five hundred different realizations were done and they were all fit by the model ***y=c*t +d***.

The cumulative distribution of both "c" and "d" were calculated, as well as the average and standard deviation of the 90% confidence limits, which should delimit the lower 5% and upper 95% of the cumulative distributions.

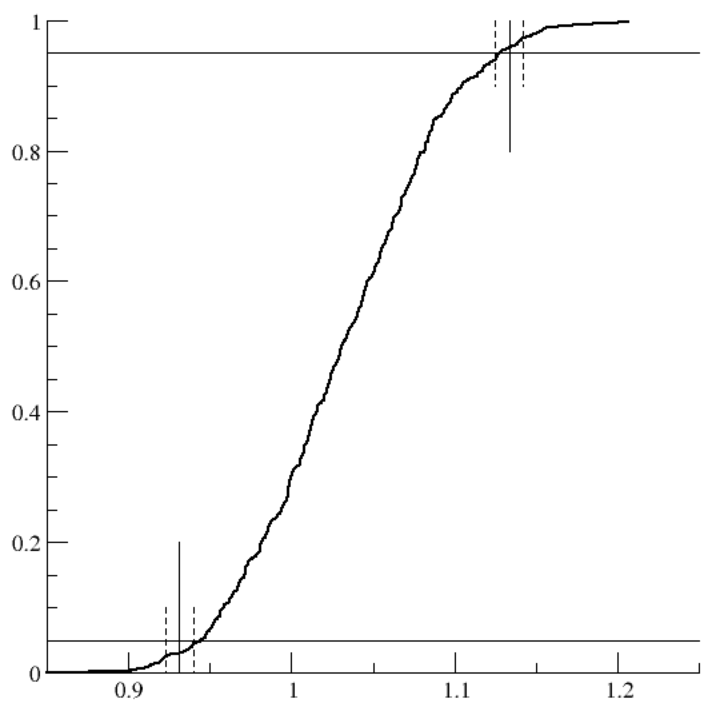

Figure 1:The cumulative distribution for "c" is drawn, the dark S shape curve in the middle of the page. The actual 5% and 95%cumulative distribution lines are drawn horizontally. The estimated 90% confidence limits (5% below, 95% above) are the larger vertical lines. The dashed lines represent plus and minus one standard deviation of the 90% confidence limits. The calculated 90% confidence limits for "c" were overestimated by approximately 10%.

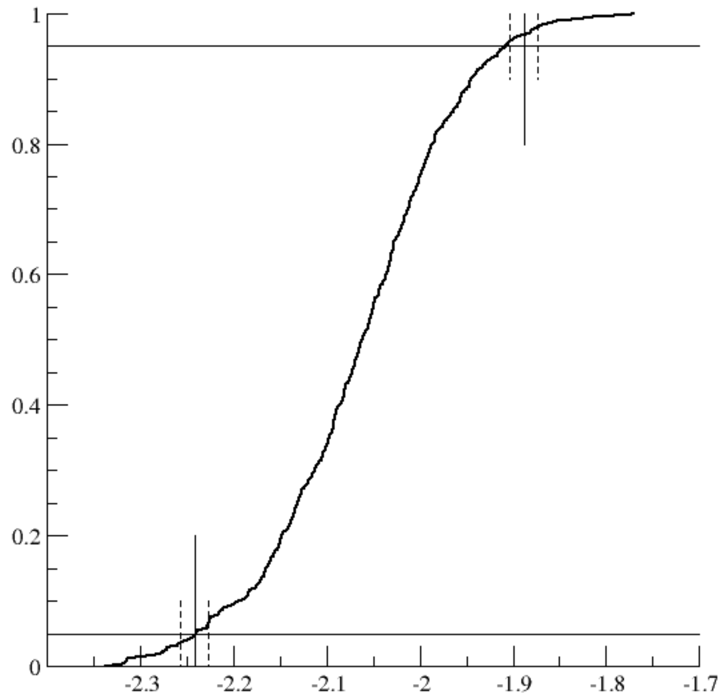

Figure 2:The cumulative distribution for “d” is drawn, the dark S shape curve in the middle of the page. The actual 5% and 95%cumulative distribution lines are drawn horizontally. The estimated 90% confidence limits (5% below, 95% above) are the larger vertical lines. The dashed lines represent plus and minus one standard deviation of the 90% confidence limits. The calculated 90% confidence limits for “d” were overestimated by approximately 5%.

We note that in the paper SSR is defined to be half it’s value. Although this makes no difference to an optimization, it does affect the calculation of the confidence interval. This appears to be an error. It would lead to an underestimation of the confidence intervals given in the two figures by approximately 50%.

~garyr/confidenceIntervals.fm